KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

# Confidence intervals for high-dimensional partially linear single-index models

Gueuning T, Claeskens G.

KBI_1611

# Confidence intervals for high-dimensional partially linear single-index models

**Thomas Gueuning and Gerda Claeskens**

ORSTAT and Leuven Statistics Research Center
KU Leuven, Faculty of Economics and Business
Naamsestraat 69, 3000 Leuven, Belgium
thomas.gueuning@kuleuven.be, gerda.claeskens@kuleuven.be

**Abstract**

We study partially linear single-index models where both model parts may contain high-dimensional variables. While the single-index part is of fixed dimension, the dimension of the linear part is allowed to grow with the sample size. Due to the addition of penalty terms to the loss function in order to provide sparse estimators, such as obtained by lasso or SCAD, the construction of confidence intervals for the model parameters is not as straightforward as in the classical low-dimensional data framework. By adding a correction term to the penalized estimator a desparsified estimator is obtained for which asymptotic normality is proven. We study the construction of confidence intervals and hypothesis tests for such models. The simulation results show that the method performs well for high-dimensional single-index models.

**Keywords**: high-dimensional data; single-index model; regularized estimation; sparsity; asymptotic normality; confidence interval

## 1 Introduction

A partially linear single-index model is a semi-parametric model which can be written as $Y = \eta(Z^\top\alpha)+X^\top\beta+\epsilon$, where $\eta$ is a one-dimensional unknown function and $Z$, $X$ are covariate vectors of dimension $p$ and $q$, respectively, such that the mean zero random error $\epsilon$ is independent of $(Z, X)$. The underlying idea is that when the linearity assumption may not be valid for all covariates the introduction of an unknown function allows to overcome this problem. We focus on the high-dimensional case where the number of covariates $p + q$ may exceed the sample size $n$. In this paper, we consider $p$ as fixed and allow $q$ to grow with $n$.

We show how to construct a desparsified version of a penalized estimator of the high-dimensional parameters $(\alpha, \beta)$ of the partially linear single-index model, and we establish the asymptotic distribution of this desparsified estimator. The main purpose of desparsifying a penalized estimator that is obtained under the assumption of sparsity is that the construction of confidence intervals becomes thus possible for *all* of the components of the parameter vector, also for the ones asymptotically consistently estimated by zero due to the penalized estimation and which result

in a point mass at zero in the asymptotic distribution. Van de Geer et al. (2014) introduce such desparsified lasso estimators in the context of linear and generalized linear models and obtain their asymptotically normal distribution. This is done by writing the Karush-Kuhn-Tucker conditions which define the lasso estimator. It is then possible to find a de-biased lasso estimator and to characterize its asymptotic distribution under suitable conditions including a stricter sparsity condition. The construction of confidence intervals for the model parameters is then straightforward. Note that for linear models, their method is the same one as that from Zhang and Zhang (2014). Javanmard and Montanari (2014) also provide a desparsified estimator in the context of the linear model. The main difference with the approach of van de Geer et al. (2014) is that they do not make any assumption on the sparsity level of the precision matrix.

Waldorp (2015) used the desparsified lasso for comparing high-dimensional graphical models. He obtained desparsified estimators for the coefficients of the precision matrices of the graphical models and constructed hypothesis tests based on the asymptotic distribution of these estimators. Lu et al. (2015) also used the desparsifying idea for constructing confidence bands for a class of nonparametric sparse additive models.

The use of a single index as opposed to a full $p$-variate nonparametric function estimation effectively circumvents the curse of dimensionality. In a low dimensional setting, the (partially linear) single-index model has been studied by Carroll et al. (1997), Horowitz (1998), Xia et al. (1999), Yu and Ruppert (2002) and Xia and Härdle (2006), among others. The fitting process involves estimation of the parameters and of the unknown function $\eta$. Different fitting methods have been introduced, most of them use kernel smoothing. Liang et al. (2010) and Wang et al. (2010) used the local linear regression technique introduced by Fan and Gijbels (1996) to estimate $\eta$. The resulting estimators have good theoretical properties regarding consistency and convergence rates. To deal with high dimensional covariate vectors, Liang et al. (2010) used a smoothly clipped absolute deviation penalty (SCAD, see Fan and Li (2001)). They obtain a profile least-squares function to minimize, similarly to the linear model case. Ma and Zhu (2013) study such models under heteroscedasticity.

Our goal is to provide confidence intervals for the high-dimensional parameter vector in the partially linear single index model and to determine which conditions (design, sparsity, etc.) are necessary to make this construction possible, the challenging part being to be able to tackle the presence of the unknown function $\eta$. In a fixed dimension framework, Zhu and Xue (2006) introduced the empirical likelihood to construct confidence regions. Further, Zhang et al. (2012) developed a dimension reduction approach for estimation in a partially linear single-index model (PLSIM) with diverging number of parameters in both the linear and the single-index part but needed the strong condition $\max(p, q) = o(n^{1/3})$ excluding the $p + q > n$ case. Our method for

constructing confidence intervals and regions works in the high-dimensional framework $p+q > n$ with $p$ fixed (potentially larger than $n$) and $q$ growing with $n$.

First, in Section 2, we describe a method for estimating a partially linear single-index model. Our main results are in Section 3 where we construct a desparsified estimator and study its asymptotic distribution. Section 4 describes the construction of confidence intervals and regions together with ways to perform hypothesis testing following from the theoretical study. Section 5 deals with computational choices. Further, Section 6 reports on simulation studies to assess the finite sample performance of the desparsified estimator. In Section 7 we illustrate our method on a dataset to study the Bardet-Biedl disease in a rat population. This disease is linked to genetic mutations and also affects humans, provoking several dysfunctions. The dataset comprises of 120 observations and 200 variables. Section 8 concludes. All proofs are contained in Section 9.

## 2 Estimation for partially linear single-index model

Let $\{(Y_i, Z_i, X_i), i = 1, \ldots, n\}$ be a sample generated by the partially linear single-index model

$$Y = \eta(Z^\top \alpha_0) + X^\top \beta_0 + \epsilon,$$

where $\eta : \mathbb{R} \to \mathbb{R}$ is an unknown differentiable function, $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, $\alpha_0 \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}^q$ are the regression parameters and $Z$ and $X$ are, respectively, $p$-dimensional and $q$-dimensional covariate vectors. The error term $\epsilon$ and the covariates $Z, X$ are independent. For identifiability reason, we assume that $\|\alpha_0\| = 1$ and that the first non-zero entry of $\alpha_0$ is positive. We consider $p$ as fixed and allow $q$ to grow with $n$. The case $p + q > n$ corresponds to the high-dimensional data framework. We denote by $\xi = (\alpha^\top, \beta^\top)^\top$ the $p + q-$ dimensional vector of parameters.

Several estimation approaches have been introduced in the literature. In this paper, we use the profile least-squares procedure presented in Liang et al. (2010). For the paper to be self-contained, we here summarize the main ingredients. This approach uses the local linear regression technique to estimate $\eta$, that is, an estimator of $\eta(u)$ is obtained by the minimization of

$$\sum_{i=1}^{n} \{a + b(Z_i^\top \alpha - u) + X_i^\top \beta - Y_i\}^2 K_h(Z_i^\top \alpha - u), \tag{1}$$

with respect to $a$ and $b$, where $K_h(\cdot) = h^{-1}K(\cdot/h)$, $K(\cdot)$ is a kernel function and $h$ is a bandwidth. The minimizer $(\hat{a}, \hat{b})$ of (1) is an estimator of $(\eta(u), d\eta(u)/du)$. It can be shown (see Fan and Gijbels (1996)) that

$$\hat{\eta}(u, \xi) = \hat{a} = \frac{K_{20}(u,\xi)K_{01}(u,\xi) - K_{10}(u,\xi)K_{11}(u,\xi)}{K_{00}(u,\xi)K_{20}(u,\xi) - K_{10}^2(u,\xi)},$$

where $K_{jl}(u, \xi) = \sum_{i=1}^n K_h(Z_i^\top \alpha - u)(Z_i^\top \alpha - u)^j (X_i^\top \beta - Y_i)^l$ for $j = 0, 1, 2$ and $l = 0, 1$.

Now, for every data point, we have an estimator $\hat{\eta}(Z_i^\top \alpha; \xi)$ of $\eta(Z_i^\top \alpha)$ and, in the low-dimensional case where $p + q < n$, we can obtain a profile least-squares estimator $\hat{\xi} = (\hat{\alpha}, \hat{\beta})$ by the minimization of

$$Q(\alpha, \beta) = \sum_{i=1}^{n} \{Y_i - \hat{\eta}(Z_i^\top \alpha; \xi) - X_i^\top \beta\}^2, \tag{2}$$

with the constraint $\|\alpha\|_2 = 1$ for identifiability reason. If the number of variables $p + q$ is large or if some sparsity is expected it can be relevant to add penalty terms to $Q(\alpha, \beta)$ so that variable selection and parameter estimation are simultaneously performed. We then consider the minimization of

$$L(\alpha, \beta) = \frac{1}{2} Q(\alpha, \beta) + n \sum_{j=1}^{p} p_{\lambda_{1j}}(|\alpha_j|) + n \sum_{k=1}^{q} p_{\lambda_{2k}}(|\beta_k|), \tag{3}$$

with the constraint $\|\alpha\|_2 = 1$, where $p_\lambda(\cdot)$ is a penalty function, such as the Lasso penalty (Tibshirani (1996)), the adaptive Lasso penalty (Zou (2006); Wang and Wu (2013)) and the SCAD penalty (Fan and Li (2001); Liang et al. (2010)). These penalty functions are defined as follows,

$$\begin{aligned}
\text{Lasso: } & p_\lambda(x) = \lambda |x| \\
\text{Adaptive Lasso: } & p_\lambda(x) = \lambda \tilde{x}^{-\gamma} |x| \\
\text{SCAD: } & p'_\lambda(x) = \lambda \left\{ I(|x| \leq \lambda) + \frac{(\nu\lambda - |x|)_+}{(\nu - 1)\lambda} I(|x| > \lambda) \right\}, \; p_\lambda(0) = 0,
\end{aligned}$$

where $\tilde{x}$ is an initial estimator, $\gamma$ is a positive number, $p'_\lambda(x)$ is the derivative of $p_\lambda(x)$ and $\nu = 3.7$.

The parameters $\lambda_{1j}$ and $\lambda_{2k}$ in (3) are the $p + q$ tuning parameters. As described in section 5, some procedures can be used to drastically reduce the number of tuning parameters to choose. Note that if we want to select only amongst the $Z$-variables and allow shrinkage for the estimators for $\alpha$ though do not wish to select amongst the $X$-variables, we can set $p_{\lambda_{2k}}(\cdot) = 0$ such that the penalized profile least-squares function to minimize becomes

$$L_Z(\alpha, \beta) = \frac{1}{2} Q(\alpha, \beta) + n \sum_{j=1}^{p} p_{\lambda_{1j}}(|\alpha_j|).$$

Alternatively, if we want to select only amongst the $X$-variables and allow shrinkage for the estimators of $\beta$ though do not wish to select amongst the $Z$-variables, we can set $p_{\lambda_{1j}}(\cdot) = 0$ and we obtain

$$L_X(\alpha, \beta) = \frac{1}{2} Q(\alpha, \beta) + n \sum_{k=1}^{q} p_{\lambda_{2k}}(|\beta_k|).$$

All the estimators presented hereabove show good asymptotic properties. With the proper conditions, the nonpenalized estimator minimizing $Q(\alpha, \beta)$ is $\sqrt{n}$-consistent and its asymptotic distribution is obtained. Regarding the penalized estimators, Liang et al. (2010) show that the use of $L(\alpha, \beta)$ with the SCAD penalty provides consistency in terms of variable selection and the asymptotic distribution of the active set of variables is obtained.

By the consistency of variable selection, *asymptotically* the true zero coefficients will be estimated to be zero, resulting in an asymptotic pointmass at zero. Conditional on the selection of the active set, the asymptotic distribution of estimators of the so-called active set is normal. The construction of confidence intervals and performing hypothesis tests are thus with this method only possible for the nonzero components of the estimator. Ignoring the coefficients that are set to zero and in particular ignoring the variability associated with the selection of the zero components may lead to incorrect inference (see also Leeb et al. (2015)). The obtention of the asymptotic distribution of the full penalized estimator vector using a desparsifying process as in van de Geer et al. (2014) is the subject of Section 3.

# 3 Desparsifying the penalized estimator

In this section, we first show how to construct a desparsified version of the penalized estimator obtained by the minimization of (3), where $p_{\lambda_{1j}}$ and $p_{\lambda_{2k}}$ can be any penalty function such as the Lasso, the adaptive Lasso or the SCAD. Secondly we describe how to construct the matrix $\hat{\Theta}$ used in the desparsifying process. Thirdly we present theoretical results if the penalized estimator is the Lasso estimator and if $\hat{\Theta}$ is constructed by the nodewise regression technique.

## 3.1 Construction of a desparsified estimator

We follow the idea of van de Geer et al. (2014) and propose the following process for the construction of a desparsified version of the penalized estimator for the parameters of a high-dimensional partially linear single-index model. We start with the penalized estimator $\hat{\xi} = (\hat{\alpha}, \hat{\beta}) = \arg\min L(\alpha, \beta)$, with $L(\alpha, \beta)$ the penalized profile least squares function (3). We use the following notations. For a parameter vector $\xi = (\alpha, \beta)$ and a triple $(y, z, x)$ we write the loss function $\rho_\xi(y, z, x) = \{y - x^\top \beta - \hat{\eta}(z^\top \alpha; \alpha, \beta)\}^2$ and we write $\rho_\xi^i = \rho_\xi(Y_i, Z_i, X_i)$ the loss function corresponding to the $i$-th sample observation $(Y_i, Z_i, X_i)$. We denote by $\dot{\rho}_\xi^i$ and $\ddot{\rho}_\xi^i$ the first and second partial derivatives of $\rho_\xi^i$ with respect to $(\alpha, \beta)$ evaluated in $\xi$. We also use the notation $P_n g = n^{-1} \sum_{i=1}^n g(Y_i, Z_i, X_i)$ and $Pg = \mathrm{E}[P_n g]$ for a general function $g$ and we write $A^{\otimes 2} = AA^\top$ for any matrix $A$.

We have the following expressions for the first and second partial derivatives of the sample mean

of the loss functions in $\hat{\xi} = (\hat{\alpha}, \hat{\beta})$,

$$\begin{aligned} P_n\dot{\rho}_{\hat{\xi}} &= -\frac{1}{n}\sum_{i=1}^{n}\{Y_i - \hat{\eta}(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta}) - X_i^\top\hat{\beta}\}\begin{bmatrix} Z_i\hat{\eta}'(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta}) \\ X_i \end{bmatrix}, \\ P_n\ddot{\rho}_{\hat{\xi}} &= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} Z_i\hat{\eta}'(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta}) \\ X_i \end{bmatrix}^{\otimes 2}. \end{aligned}$$

We define $\hat{\Sigma} = P_n\ddot{\rho}_{\hat{\xi}}$ and construct a "relaxed" inverse $\hat{\Theta}$ of $\hat{\Sigma}$ (see Section 3.2). We now define the following desparsified estimator

$$\hat{\xi}^{\text{desp}} = \hat{\xi} - \hat{\Theta}P_n\dot{\rho}_{\hat{\xi}}. \tag{4}$$

The construction of this desparsified estimator for partially linear single-index models is similar to what is done in van de Geer et al. (2014) for (generalized) linear models and despite the presence of nonparametric effects we are able to track the asymptotic distribution of this desparsified estimator. The term $\hat{\Theta}P_n\dot{\rho}_{\hat{\xi}}$ can be seen as a bias correction term to the Lasso estimator that prevents the asymptotic distribution to contain a pointmass at zero.

### 3.2 Construction of the relaxed inverse $\hat{\Theta}$ of $\hat{\Sigma}$

There are several techniques to construct a relaxed inverse of $\hat{\Sigma}$. We present here two of them, introduced respectively by Javanmard and Montanari (2014) and van de Geer et al. (2014). These two techniques lead to two different desparsified estimators but in practice we observe that they tend to be close to each other.

#### Convex optimization

Let us denote by $\hat{\Theta}_{j\bullet}$ the $j$th row of $\hat{\Theta}$ and let us consider it as a $(p+q)\times 1$ vector. Using a Taylor expansion of $P_n\dot{\rho}_\xi$ around $\xi_0$ we can show the following,

$$\hat{\xi}_j^{\text{desp}} - \xi_j^0 = -\hat{\Theta}_{j\bullet}^\top P_n\dot{\rho}_{\xi_0} - (\hat{\Theta}_{j\bullet}^\top\hat{\Sigma} - e_j^\top)(\hat{\xi}-\xi_0) + \hat{\Theta}_{j\bullet}^\top P_n(\ddot{\rho}_{\hat{\xi}} - \ddot{\rho}_{\bar{\xi}})(\hat{\xi}-\xi_0) \equiv T_j + r_1 + r_2,$$

where $\bar{\xi}$ is an interior point between $\xi_0$ and $\hat{\xi}$ and $e_j \in \mathbb{R}^{p+q}$ is the $j$-th unit vector consisting of a one at the $j$ component and zeros elsewhere. We expect $\sqrt{n}T_j$ to be asymptotically normally distributed, with its variance determined by the limit of $\hat{\Theta}_{j\bullet}^\top\hat{\Sigma}\hat{\Theta}_{j\bullet}$. This will be formalized in Theorem 1. We also expect $r_2$ to be small because the difference between $\ddot{\rho}_{\bar{\xi}}$ and $\ddot{\rho}_{\hat{\xi}}$ tends to zero. Thus, the matrix $\hat{\Theta}$ should be chosen so that the remainder term $r_1$ and the variance term $\hat{\Theta}_{j\bullet}^\top\hat{\Sigma}\hat{\Theta}_{j\bullet}$ are small. This leads to the following process introduced by Javanmard and Montanari (2014): for $j \in \{1,\ldots,p+q\}$, let $\hat{\Theta}_{j\bullet}$ be the solution of the following convex problem,

$$\text{minimize } \Theta_{j\bullet}^\top\hat{\Sigma}\Theta_{j\bullet} \text{ with respect to } \Theta_{j\bullet} \in \mathbb{R}^{p+q}, \text{ subject to } \|\hat{\Sigma}\Theta_{j\bullet} - e_j\|_\infty \le \kappa,$$

where $\kappa$ is a tuning parameter. If some of the convex minimizations are not feasible then $\hat{\Theta} = I_{p+q,p+q}$. We refer to Javanmard and Montanari (2014) for the theoretical properties of this estimator and its use for debiasing the lasso estimator.

**Nodewise regression**

The second technique, introduced by Meinshausen and Buhlmann (2006) and used by van de Geer et al. (2014), assumes that the matrix $\Theta = \Theta_{\xi_0}$ which serves as an approximation to what would be the inverse matrix of $P\ddot{\rho}_{\xi_0}$ (which might not exist in the high-dimensional case) is sparse. The process presented below guarantees the sparsity of $\hat{\Theta}$ and allows to control $\|\hat{\Theta}_j^\top \hat{\Sigma} - e_j^\top\|_\infty$ through the tuning parameters $\lambda_j$. More details can be found in van de Geer et al. (2014).
For the partially linear single index models, define a new $n \times (p+q)$ design matrix $D_{\hat{\xi},\hat{\eta}}$ of which the $i$th row is given by

$$D_{(\hat{\xi},\hat{\eta}),i\bullet} = (Z_i^\top \hat{\eta}'(Z_i^\top \hat{\alpha}; \hat{\xi}), X_i^\top),$$

for $i = 1, \ldots, n$. We thus have $P_n \ddot{\rho}_{\hat{\xi}} = n^{-1} \sum_{i=1}^n D_{(\hat{\xi},\hat{\eta}),i\bullet} D_{(\hat{\xi},\hat{\eta}),i\bullet}^\top$ for which we want to find an approximated inverse. To motivate the approach, if the number of parameters would be smaller than the sample size, the matrix $P_n \ddot{\rho}_{\hat{\xi}}^{-1}$ would be the inverse of the variance matrix of the least squares regression coefficient estimator in a linear regression model which takes $D_{(\hat{\xi},\hat{\eta}),i\bullet}$ as the covariate vector for the $i$th observation and $Y_i$ as the response, for $i = 1, \ldots, n$. This suggests that in the high-dimensional setting to estimate a relaxed inverse of $P_n \ddot{\rho}_{\hat{\xi}} = n^{-1} D_{(\hat{\xi},\hat{\eta})} D_{(\hat{\xi},\hat{\eta})}^\top$ we can use a nodewise penalized regression of the $j$th column of $D_{(\hat{\xi},\hat{\eta})}$, denoted by $D_{(\hat{\xi},\hat{\eta}),\bullet j}$, on all other columns, denoted by $D_{(\hat{\xi},\hat{\eta}),\setminus j}$. Next, we proceed in a similar way as in van de Geer et al. (2014) and construct for each $j = 1, \ldots, p+q$ the lasso estimator

$$\hat{\gamma}_j = \underset{\gamma \in \mathbb{R}^{p+q-1}}{\arg\min} \{\|D_{(\hat{\xi},\hat{\eta}),\bullet j} - D_{(\hat{\xi},\hat{\eta}),\setminus j}\gamma\|_2^2/n + 2\lambda_j \|\gamma\|_1\}, \tag{5}$$

of which the components are denoted by $\hat{\gamma}_{j,k}$ with $k \neq j$, $k = 1, \ldots, p+q$. Equation (5) corresponds to equations (7) and (19) of van de Geer et al. (2014) for linear models and generalized linear models. With $\hat{\tau}_j^2 = (D_{(\hat{\xi},\hat{\eta}),\bullet j} - D_{(\hat{\xi},\hat{\eta}),\setminus j}\hat{\gamma}_j)^\top D_{(\hat{\xi},\hat{\eta}),\bullet j}/n$, we define the relaxed inverse $\hat{\Theta}$ by

$$\hat{\Theta} = \operatorname{diag}\left(\hat{\tau}_1^{-2}, \ldots, \hat{\tau}_{p+q}^{-2}\right) \cdot \begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \ldots & -\hat{\gamma}_{1,p+q} \\ -\hat{\gamma}_{2,1} & 1 & \ldots & -\hat{\gamma}_{2,p+q} \\ \vdots & \vdots & \ddots & \vdots \\ -\hat{\gamma}_{p+q,1} & -\hat{\gamma}_{p+q,2} & \ldots & 1 \end{pmatrix}.$$

### 3.3 Theoretical results

Recall that we write $\xi = (\alpha, \beta)$ the $p+q$ dimensional vector of parameters with $\xi_0$ the true value. We define $g_{2i}(t)$ as the $i$th component of $g_2(t) = \mathrm{E}[X|Z^\top\alpha = t]$, $1 \leq i \leq q$ and $g_{3j}(t)$ is the $j$th component of $g_3(t) = \mathrm{E}[Z|Z^\top\alpha = t]$, $1 \leq j \leq p$.
We consider the penalized estimator $\hat{\xi}$ as the solution of the optimization problem

$$\min_{\xi=(\alpha,\beta)\in\mathbb{R}^{p+q}} \frac{1}{2n}\sum_{i=1}^{n}\{Y_i - \hat{\eta}(Z_i^\top\alpha;\xi) - X_i^\top\beta\}^2 + p_\lambda(\xi). \tag{6}$$

Before listing the conditions needed for our theoretical results, we give some definitions. The parameter $\lambda$ is the tuning parameter in (6) and the parameter $\lambda_*$ is such that for all $j$, $\lambda_* \asymp \lambda_j$ where $\lambda_j$ is the tuning parameter used in (5). Recall that we consider $p$ as fixed and $q$ as being allowed to grow with $n$. The parameter $s_0$ is the sparsity level of the model defined as the number of non-zero components of $\xi_0$. Recall that $h$ is the bandwidth used to estimate the function $\eta$. For the estimation of the derivative $\eta'$, the use of the same bandwidth $h$ would lead to a slower convergence rate than that of $\eta$. In order to control this convergence rate, we decide to use another bandwidth $h_1$ for the estimation of the derivative, as in Wang et al. (2010). We now list the following conditions in order to establish the asymptotic normality of the desparsified estimator.

(C1) (i) The distribution of $Z$ has a compact support set $A$;

(ii) The density function of $Z^\top\alpha$ is positive and satisfies a Lipschitz condition of order 1 for $\alpha$ in a neighborhood of $\alpha_0$. The density of $Z^\top\alpha_0$ is bounded on $T = \{t = z^\top\alpha : z \in A\}$.

(C2) The functions $\eta$ and $g_{2i}$ have two bounded and continuous derivatives and the function $g_{3j}$ satisfies a Lipschitz condition of order 1.

(C3) The kernel $K$ is a bounded, continuous and symmetric probability density function, satisfying $\int_{-\infty}^{\infty} u^2K(u)du \neq 0$ and $\int_{-\infty}^{\infty} |u|K(u)du < \infty$. Furthermore, $K$ satisfies a Lipschitz condition.

(C4) $nhh_1^3/\ln^2 n \to \infty$, $nh^4 \to 0$ and $\limsup_{n\to\infty} nh_1^5 < \infty$.

(C5) $\sup_t \mathrm{E}[\|X\|^2 |Z^\top\alpha_0 = t] < \infty$, $\mathrm{E}[\epsilon] = 0$, $\mathrm{Var}[\epsilon] = \sigma_\epsilon^2 < \infty$ and $\mathrm{E}[\epsilon^4] < \infty$.

(C6) (i) $\hat{\alpha} - \alpha_0 = O_P(n^{-1/2})$, (ii) $\frac{1}{n}\sum_{i=1}^{n} X_i^\top(\hat{\beta} - \beta_0) = o_P(n^{-1/2})$ and (iii) $\|\hat{\xi} - \xi_0\|_1 = O_P(\lambda s_0)$.

(C7) $\sup_{i,j} |Z_i^\top\Theta_{j,1}| = O_P(1)$ and $\sup_{i,j} |X_i^\top\Theta_{j,2}| = O_P(1)$.

(C8) $\left\| e_j^\top - \hat{\Theta}_j^\top \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} Z_i \hat{\eta}'(Z_i^\top \hat{\alpha}; \hat{\alpha}, \hat{\beta}) \\ X_i \end{bmatrix}^{\otimes 2} \right\| = O_P(\lambda_*)$.

(C9) $s_0 \lambda \lambda_* = o(n^{-1/2})$.

Conditions (C1) to (C5) are also used in Wang et al. (2010). Condition (C4) determines the range of the bandwidths $h$ and $h_1$ while the other conditions are standard for local linear smoothing. Condition (C6)(i) is an assumption on the consistency of the low dimensional estimator $\hat{\alpha}$ justified by the theoretical results of Liang et al. (2010) and Wang et al. (2010), while (C6)(iii) is a usual property for a high-dimensional estimator. (C7) and (C8) are conditions on the relaxed inverse $\hat{\Theta}$ and (C9) is a sparsity assumption. With $\lambda = O(\sqrt{\ln(p+q)/n})$ and $\lambda_* = O(\sqrt{\ln(p+q)/n})$, (C9) holds if we have $s_0 = o(\sqrt{n}/\ln(p+q))$. Conditions similar to (C7) to (C9) are used in van de Geer et al. (2014).

**Theorem 1.** *Let $\hat{\xi}^{\text{desp}}$ be obtained by the desparsifying process (4) where $\hat{\Theta}$ is obtained by the nodewise regression technique. Under conditions (C1)-(C9), we have:*

$$\sqrt{n}(\hat{\xi}^{\text{desp}} - \xi_0) = A + o_P(1),$$

*with $A \to_d \mathcal{N}(0, \Sigma_A)$. It holds that*

$$S_{A,n} = \sigma_\epsilon^2 \hat{\Theta} \frac{1}{n} \sum_{i=1}^n \begin{bmatrix} \tilde{Z}_i \hat{\eta}'(Z_i^\top \hat{\alpha}; \hat{\xi}) \\ \tilde{X}_i \end{bmatrix}^{\otimes 2} \hat{\Theta}^\top$$

*is a consistent estimator of $\Sigma_A$ with $\tilde{Z}_i = Z_i - \mathrm{E}[Z|Z_i^\top \alpha_0]$ and $\tilde{X}_i = X_i - \mathrm{E}[X|Z_i^\top \alpha_0]$.*

The proof can be found in section 9.

**Remark 1.** *The distribution of any sub-vector $\hat{\xi}_L^{\text{desp}}$ of $\hat{\xi}^{\text{desp}}$ can be readily obtained by Theorem 1. It has mean $\xi_{0,L}$ corresponding to the indices in $L$ and the variance is estimated by replacing $\hat{\Theta}$ by $\hat{\Theta}_L$, the $|L| \times p + q$ sub-matrix of $\hat{\Theta}$ corresponding to the indices in $L$.*

**Remark 2.** *The reason to consider the dimension $p$ of $\alpha_0$ as being fixed while the dimension $q$ of $\beta_0$ is authorized to grow with $n$ is that Theorem 1 requires the use of the Lemma A.4 from Wang et al. (2010) which holds only for fixed $p$. Indeed, the proof of this lemma involves a covering number $n^{2pa}$ that needs to be dominated by $n^{-cM^2}$ for a constant $M$ large enough, with $a$ and $c$ positive constants.*

## 4 Construction of confidence intervals and hypothesis testing

Based on Theorem 1, we construct univariate confidence intervals and multivariate confidence regions as follows. Let us denote by $\hat{\sigma}_j^2 = S_{A,n;j,j}$ the $j$-th element of the diagonal of $S_{A,n}$. If $\sigma_\epsilon$

is unknown, we replace it by a consistent estimator. A confidence interval at a confidence level $1-\alpha$ for a component $\xi_{0,j}$ of the true parameter is defined as

$$\mathrm{CI}_j = \left[\hat{\xi}_j^{\mathrm{desp}} - \frac{\hat{\sigma}_j}{\sqrt{n}}\Phi^{-1}(1-\alpha/2), \hat{\xi}_j^{\mathrm{desp}} + \frac{\hat{\sigma}_j}{\sqrt{n}}\Phi^{-1}(1-\alpha/2)\right], \tag{7}$$

where $\hat{\xi}_j^{\mathrm{desp}}$ is defined in Theorem 1 and where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Furthermore, for a finite set $L \subset \{1,\ldots,p+q\}$ we define a confidence region at a confidence level $1-\alpha$ as

$$\mathrm{CR}_L = \left\{\xi_L \in \mathbb{R}^{|L|} : n(\xi_L - \hat{\xi}_L^{\mathrm{desp}})^\top S^{-1}_{A,n;L\times L}(\xi_L - \hat{\xi}_L^{\mathrm{desp}}) \leq q\chi^2_{|L|}(1-\alpha)\right\}. \tag{8}$$

with $q\chi^2_{|L|}(1-\alpha)$ the $1-\alpha$ quantile of the chi-square distribution with $|L|$ degrees of freedom. As pointed out by Leeb et al. (2015), Kabaila (2009) and Hjort and Claeskens (2003), amongst others, it is still quite common amongst practitioners to follow a model selection procedure by inference conditionally on the selected model. However, this leads to inaccurate inference because it does not take the uncertainty associated with the model selection process into account. In a penalized estimation framework, this means that constructing confidence intervals or regions based on the asymptotic distribution of the active set of components, using the oracle property, might be misleading and should be avoided. In particular for the parameters of a partially linear single-index model, one should avoid to construct confidence intervals based on normality results of the estimated active set of coefficients only, such as Theorem 2 in Liang et al. (2010). Instead, we propose to use constructions based on a desparsifying process such as the confidence interval (7) and the confidence region (8) in order to correctly take the uncertainty associated with the variable selection process into account. This approach based on the desparsified lasso is thus different from concentrating on the active set of a lasso path as studied by Lockhart et al. (2014).

Theorem 1 can also be used for testing significance of variables as follows. For a component $j \in \{1,\ldots,p+q\}$, we test

$$H_0^1 : \xi_{0,j} = 0 \text{ versus } H_a^1 : \xi_{0,j} \neq 0$$

and under $H_0^1$, we have $|\sqrt{n}\hat{\xi}_j^{\mathrm{desp}}/\hat{\sigma}_j| \leq \Phi^{-1}(1-\alpha/2)$ with probability $1-\alpha$. We also test simultaneous significance as follows. For a finite set $L \subset \{1,\ldots,p+q\}$ we test

$$H_0^2 : \xi_{0,j} = 0 \text{ for all components } j \in L \text{ versus } H_a^2 : \xi_{0,j} \neq 0 \text{ for at least one component } j \in L$$

and, under $H_0^2$, we have $n\hat{\xi}_L^{\mathrm{desp}\top} S^{-1}_{L\times L}\hat{\xi}_L^{\mathrm{desp}} \leq q\chi^2_{|L|}(1-\alpha)$ with probability $1-\alpha$.

As a conservative alternative to using the joint asymptotic distribution of the estimated components, one could compute a confidence interval $\mathrm{CI}_j$ for every component $j$ of $L$ and then apply

a Bonferroni correction to control the family wise error rate. The use of the joint distribution has as an additional benefit that also the correlations between the components of $\hat{\xi}_L^{\text{desp}}$ is taken into account.

# 5 Computational choices

We briefly describe how to choose the bandwidths and the tuning parameters $\lambda_{1j}$ and $\lambda_{2k}$ in (3). We also describe how to make the minimization of (2) and (3) faster by using a local linear approximation.

## 5.1 Choice of the bandwidths

There are many ways to choose a bandwidth $h$ for the local linear regression. A rule of thumb given by Scott (2009) is to use $h_{\text{ROT}} = 2.756\hat{\sigma}_\epsilon n^{-1/5}$ where $\hat{\sigma}_\epsilon$ is an estimator of the standard deviation of the noise. Another possibility is to select the bandwidth $h_{\text{GCV}}$ that minimizes the generalized cross validation criterion

$$\text{GCV}(h) = n^{-1}\sum_{i=1}^{n}\{Y_i - X_i^\top\hat{\beta} - \hat{\eta}_h(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta})\}^2/\{n^{-1}\text{tr}(I_n - S_h)\}^2$$

with $S_h$ the $n \times n$ smoothing matrix corresponding to $h$. These two methods are expected to produce bandwidths of order $O(n^{-1/5})$. When using two different bandwidths for $\hat{\eta}$ and $\hat{\eta}'$, we can choose $h_1$ as $h_{\text{GCV}}$ or $h_{\text{ROT}}$, and $h = h_1 n^{1/5} n^{-1/3} = h_1 n^{-2/15}$ (see Carroll et al. (1997) and Wang et al. (2010)) in order to satisfy the condition (C4) of theorem 1. In practice, with data of finite sample size it is not clear which strategy is the best one and numerical results (not shown here) suggest that it is acceptable to work with $h = h_1 = h_{\text{ROT}}$.

## 5.2 Choice of the tuning parameters

A common method to choose the $p + q$ tuning parameters in (3) is first to reduce their number to a more manageable number and then to perform a grid search with the BIC as selecting criterion.

For the adaptive Lasso, Huang et al. (2008) suggest to set $\lambda_{1j} = \lambda$, $\lambda_{2k} = \tau$ and $\gamma$ in $\{0.5, 1, 2\}$ and then to select the triple $\Gamma = (\gamma, \lambda, \tau)$ by minimizing

$$\text{BIC}(\Gamma) = \ln\{\text{MSE}(\Gamma)\} + n^{-1}\ln(n)\text{df}(\Gamma)$$

where $\text{MSE}(\Gamma) = n^{-1}\sum_{i=1}^{n}\{Y_i - \hat{\eta}(Z_i^\top\hat{\alpha}_\Gamma;\hat{\alpha}_\Gamma,\hat{\beta}_\Gamma) - X_i^\top\hat{\beta}_\Gamma\}^2$ and $\text{df}(\Gamma)$ is the number of nonzero coefficients of $\hat{\alpha}_\Gamma$ and $\hat{\beta}_\Gamma$ combined, the minimizers of the penalized profile least-squares function. A similar expression can be found in Zhang et al. (2012).

For the SCAD penalty, Liang et al. (2010) suggest to set $\lambda_{1j} = \lambda \mathrm{SE}(\hat{\alpha}_j^u)$ and $\lambda_{2k} = \lambda \mathrm{SE}(\hat{\beta}_k^u)$ where $\mathrm{SE}(\hat{\alpha}_j^u)$ and $\mathrm{SE}(\hat{\beta}_k^u)$ are the standard errors of the unpenalized estimators $\hat{\alpha}_j$ and $\hat{\beta}_k$ obtained by the minimization of (2) and then to select $\lambda$ using the BIC selector.

### 5.3 Linearization

In Liang et al. (2010), the Newton-Raphson algorithm is used to obtain the estimator $\hat{\xi}$ minimizing $Q(\alpha, \beta)$ or a penalized estimator minimizing $L(\alpha, \beta)$. If the dimension $p+q$ of the parameters is large and if a penalty term is used, this procedure can be computationally complex. In such case it can be interesting to use the procedure introduced in Wang and Wu (2013) which consists of using a local linear approximation of $\hat{\eta}(Z_i^\top \alpha; \alpha, \beta)$ around an initial value $(\tilde{\alpha}, \tilde{\beta})$ resulting in the one-step estimator

$$\hat{\eta}(Z_i^\top \alpha; \alpha, \beta) = \hat{\eta}(Z_i^\top \tilde{\alpha}; \tilde{\alpha}, \tilde{\beta}) + \frac{\partial \hat{\eta}}{\partial(\alpha, \beta)}|_{(\tilde{\alpha}, \tilde{\beta})} \begin{pmatrix} \alpha - \tilde{\alpha} \\ \beta - \tilde{\beta} \end{pmatrix}. \tag{9}$$

Then $Q(\alpha, \beta)$ can be approximated by

$$Q_L(\alpha, \beta) = \sum_{i=1}^{n} \left\{ Y_i - \hat{\eta}(Z_i^\top \tilde{\alpha}; \tilde{\alpha}, \tilde{\beta}) - \frac{\partial \hat{\eta}}{\partial(\alpha, \beta)}|_{(\tilde{\alpha}, \tilde{\beta})} \begin{pmatrix} \alpha - \tilde{\alpha} \\ \beta - \tilde{\beta} \end{pmatrix} - X_i^\top \beta \right\}^2.$$

By letting

$$Y_i^* = Y_i - \hat{\eta}(Z_i^\top \tilde{\alpha}; \tilde{\alpha}, \tilde{\beta}) + \frac{\partial \hat{\eta}}{\partial(\alpha, \beta)}|_{(\tilde{\alpha}, \tilde{\beta})} \begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix}, \quad X_i^* = \frac{\partial \hat{\eta}}{\partial(\alpha, \beta)}|_{(\tilde{\alpha}, \tilde{\beta})} + (\mathbf{0}_{1\times p}, X_i^\top),$$

we obtain

$$Q_L(\alpha, \beta) = \sum_{i=1}^{n} \left\{ Y_i^* - X_i^* \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \right\}^2.$$

This ordinary least squares formulation makes the computation quite efficient. Furthermore it is easy to introduce classical penalty functions like the Lasso or the SCAD and to use powerful algorithms such as the LARS algorithm introduced by Efron et al. (2004). For example, we can use the adaptive Lasso, as suggested by Wang and Wu (2013), as follows. Let us define $\zeta_{\alpha_j} = |\tilde{\alpha}_j|^{-\gamma}$ and $\zeta_{\beta_j} = |\tilde{\beta}_j|^{-\gamma}$, with $\gamma$ a positive number. An estimator of $(\alpha, \beta)$ can be obtained by minimizing the function

$$\mathcal{L}(\alpha, \beta) = Q_L(\alpha, \beta) + \lambda_n \sum_{i=1}^{p} \zeta_{\alpha_j} |\alpha_j| + \tau_n \sum_{i=1}^{q} \zeta_{\beta_j} |\beta_j|$$

with the constraint $\|\alpha\|_2 = 1$. If the initial value $(\tilde{\alpha}, \tilde{\beta})$ is $\sqrt{n}$-consistent, the asymptotic distribution of the estimator minimizing $Q_L(\alpha, \beta)$ has been obtained by Wang and Wu (2013). The

same kind of results are gotten by Wang and Wu (2013), using $\mathcal{L}(\alpha, \beta)$ with the adaptive Lasso. In practice, there are several possibilities for chosing the initial estimate $(\tilde{\alpha}, \tilde{\beta})$. In case of low-dimensional data, the least squares estimate can be used while in case of high-dimensional data we can use the lasso estimate as an initial estimate (as in Chatterjee and Lahiri (2013)) or the marginal regressor (as in Huang et al. (2008)) $\tilde{\alpha}_k = z_k^\top y/n$ and $\tilde{\beta}_k = x_k^\top y/n$ where $z$, $x$ and $y$ are the standardized $Z$, $X$ and $Y$. According to the numerical studies in Wang and Wu (2013), the procedure is fairly robust to the choice of the initial value $(\tilde{\alpha}, \tilde{\beta})$.

# 6 Simulation studies

We perform simulation studies in order to confirm our theoretical findings and to assess the finite sample behaviour.

## 6.1 Methods and models

We consider the model

$$Y = (Z^\top \alpha_0 - 0.5)^2 + X^\top \beta_0 + \sigma_\epsilon \epsilon$$

where $\epsilon$ is from a $\mathcal{N}(0,1)-$distribution and where $Z$ is independent of $X$. We consider different settings for our simulations, characterized by the following elements: (i) $n$, the number of observations: 100,200, or 500; (ii) $p$, the dimension of $\alpha_0$: 10, 50 or 200; (iii) $q$, the dimension of $\beta_0$, which is set equal to $p$ in this simulation study; (iv) $\tilde{s}_0$, the number of nonzero elements of $\alpha_0$ and $\beta_0$: 2, 5, 10, 15, 20; (v) $\sigma_\epsilon$, the noise level : 0.3 or 1; (vi) covariance matrices of $Z$ and $X$: independent, Toeplitz ($\Sigma_{j,k} = 0.9^{|j-k|}$) or equicorrelated ($\Sigma_{j,k} = 0.8$). The design matrices $Z$ and $X$ are both generated from $\mathcal{N}_p(0, \Sigma)$. Note that the sparsity level of the model is $s_0 = 2\tilde{s}_0$. All of our results are based on 1000 simulations, in which $\epsilon$, $Z$ and $X$ are randomly generated. The parameters $\alpha_0$ and $\beta_0$ are defined by $p$, $q$ and $\tilde{s}_0$ as follows. First, define $1_r$ and $0_r$ the $r \times 1$ vectors consisting of, respectively, all ones and all zeros. Then $\alpha_0 = (1/\sqrt{\tilde{s}_0} \cdot 1_{\tilde{s}_0}^\top, 0_{p-\tilde{s}_0}^\top)^\top$ and $\beta_0 = (1_{\tilde{s}_0}^\top, 0_{q-\tilde{s}_0}^\top)^\top$ where $\|\alpha_0\| = 1$ for identifiability reasons. For each simulation, we use the local linear approximation (2) to which we add a penalty function. We apply the scaled Lasso (see Sun and Zhang (2012)) which consists in using the classical Lasso with the tuning parameter $\lambda$ chosen by a data driven iterative procedure. This procedure aims at selecting a penalty level proportional to the noise level. The first choice is the scaled Lasso while results for SCAD and adaptive Lasso are also shown for some cases. Selection of the bandwidth $h$ is made by using the rule of thumb $h = 2.756\hat{\sigma}_\epsilon n^{-1/5}$ with boundary adjustment and with $\hat{\sigma}_\epsilon^2 = \mathrm{RSS}/(n - \widehat{\mathrm{df}})$. For this simulation study we use the same bandwidth $h_1 = h$ for the estimation of $\eta'$. For the estimation for a single dataset, a refined choice of the bandwidths would be possible. Additional simulation

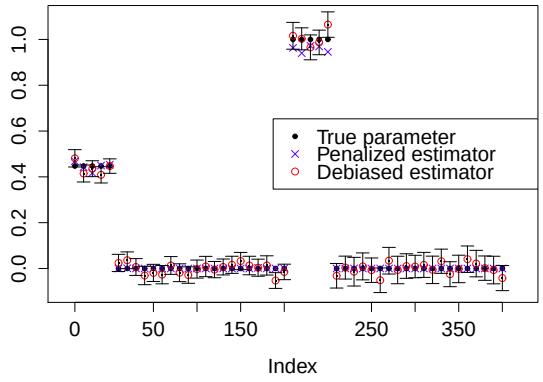


Figure 1: Simulation study. 95% confidence intervals for one realization with $(n, p, q, \tilde{s}_0) = (500, 200, 200, 5)$. For clarity only 10% of the parameters is shown.

results (not shown here) confirm that $h_1 = h$ gives good results. We thus obtain a penalized estimator $(\hat{\alpha}, \hat{\beta})$. In case of low-dimensional data we also compute the non-penalized estimator $(\hat{\alpha}, \hat{\beta})_{\text{NonPen}}$. We then construct a desparsified estimator as defined in (4), see Section 3 which we use next to construct univariate confidence intervals and multivariate confidence regions, as implied by Theorem 1. For comparison purposes we also construct univariate confidence intervals for the non-penalized estimator and for the active set of the penalized estimator. Using theoretical results from Liang et al. (2010) and Wang and Wu (2013), we construct confidence intervals and regions by directly using the non-penalized and penalized estimators.

### 6.2 Univariate confidence intervals

Using Theorem 1, we define the confidence interval for a component $\xi_{0,j}$ at a confidence level $\alpha$ as in (7). Figure 1 shows an example of univariate confidence intervals for one realization.
For each setting, we compare the empirical average coverages and the lengths of the three ways to construct a confidence interval, i.e., by the use of the non-penalized, the penalized and the desparsified estimator. Writing $\xi_0 = (\alpha_0, \beta_0)$ and $\text{CI}_j$ the confidence interval for the $j$th component of $\xi_0$, we define the empirical average coverage and length of a set $C \subset \{1, \ldots, p+q\}$ as

$$\text{Avg cov}(C) = |C|^{-1} \sum_{j \in C} \Pr(\xi_{0,j} \in \text{CI}_j) \text{ and Avg length}(C) = |C|^{-1} \sum_{j \in C} \text{E}[\text{length}(\text{CI}_j)].$$

Particular sets of interest are:

(a) $C = \{1, \ldots, p+q\}$ the full set of components;

(b) $C_{0,\alpha}$ the active set of $\alpha$, of length $\tilde{s}_0$;

(c) $C_{0,\alpha}^c$ the nonactive set of $\alpha$, of length $p - \tilde{s}_0$;

(d) $C_{0,\beta}$ the active set of $\beta$, of length $\tilde{s}_0$;

(e) $C_{0,\beta}^c$ the nonactive set of $\beta$, of length $q - \tilde{s}_0$.

We always consider 95% confidence intervals. Table 1 shows the results for different values of $n$ and of $(p, q)$. We see that the desparsified estimator provides by far the best results in every setting, with empirical coverages often close to 0.95 if $n$ is large enough. Even when the number of observations is only 100, the results are satisfying.

In Figure 2, we challenge the sparsity conditions and observe interesting results. When $\tilde{s}_0$ is large, the desparsifying estimator provides too large confidence intervals, leading to observed coverages of about 0.98 for $\tilde{s}_0 = 40$, while the non-penalized estimator provides a coverage close to 0.75. The reason is that for such large value of $\tilde{s}_0$, it does not make sense anymore to use a penalized estimator.

In Table 2, we can see the results for the three different choices of the covariance matrices and for two values of $\sigma_\epsilon$, with $(n, p, q, \tilde{s}_0)$ being set to $(500, 50, 50, 5)$. We see that the empirical coverage tends to be larger than 0.95 if the variables are correlated. We can explain this phenomenon as follows. The length of a confidence interval $\text{CI}_j$ is $2\hat{\sigma}_j \phi(1 - \alpha/2)/\sqrt{n}$ where $\hat{\sigma}_j$ is proportional to $\hat{\sigma}_\epsilon$. Thus, an overestimation of $\hat{\sigma}_\epsilon$ leads to too large confidence intervals. We also observe that the results are better when $\sigma_\epsilon$ is larger. This is also related to the estimator $\hat{\sigma}_\epsilon$. We can observe that if the noise is too small, the ratio $\sigma_\epsilon/\hat{\sigma}$ is further away from 1, leading to wrong confidence interval lengths. So, even if the estimator $\hat{\xi}$ is better, the coverage can be worse. Of course, if the noise is really large, the estimator $\hat{\xi}$ becomes so bad that the confidence intervals are not centered at the right location, leading to poor coverage. This explains why, in Table 2, the results are better for $\sigma_\epsilon = 1$ than for $\sigma_\epsilon = 0.3$.

Table 3 shows the results for three different penalty functions: SCAD, adaptive Lasso and scaled Lasso. We see that their desparsified versions provide equally good results.

To investigate a possible effect of the magnitude of the coefficients, we also performed a numerical study with components of $\alpha_0$ and $\beta_0$ of different magnitudes. In particular, we considered $\beta_0 = (3, 2, 1, 0.5, 0.2, 0, \ldots, 0)$ and $\alpha_0 = \beta_0 / \|\beta_0\|$. We did not observe any significant difference in term of average coverage between the different components of the active set. Even if a component is wrongly set to zero in the penalization procedure, the confidence interval may contain the true value of the component. The settings above have the active set of $\alpha_0$ and $\beta_0$ as $\{1, \ldots, s_0\}$ but we have also performed a numerical study with a shuffled support, that is, not all non-zero coefficients appear at the start of the vectors. We observed that shuffling the support

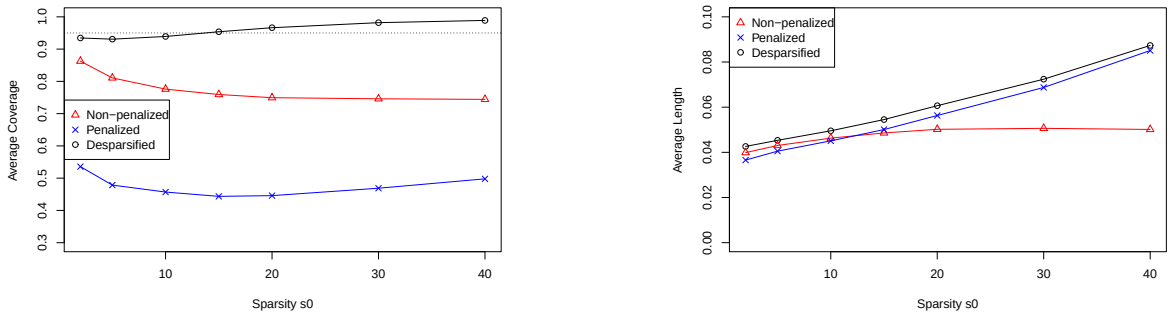


Figure 2: Average coverage and length of 95% confidence intervals in function of the sparsity parameter $\tilde{s}_0$ over 1000 realizations with $n = 500$ and $p, q = 50$.

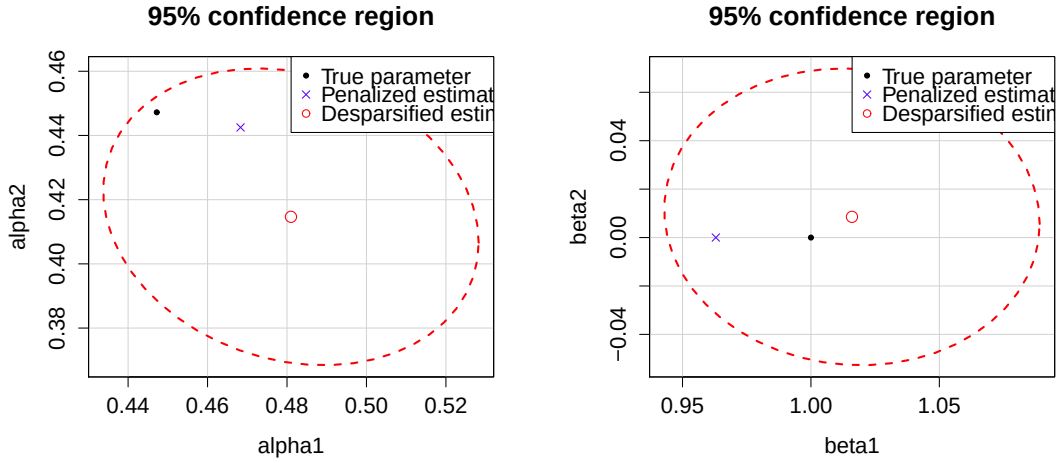


Figure 3: Examples of 2-dimensional confidence regions for components of parameters $\alpha$ and $\beta$.

makes things easier in the Toeplitz case and does not change anything in the equicorrelated and the independent case. This result is in line with van de Geer et al. (2014) (tables 1 to 4).

### 6.3 Multivariate confidence regions

We now use the multivariate version of Theorem 1 to construct confidence regions as in (8). We test our method with $n = 500$, $(p, q) = (10, 10)$ or $(50, 50)$, $\tilde{s}_0 = 2$ or 5, $\sigma_\epsilon = 0.3$ and independent or Toeplitz correlated variables. We compute over 1000 simulation runs confidence regions and average coverage for the following sets of components, where $\alpha_1, \ldots, \alpha_{\tilde{s}_0}$ and $\beta_1, \ldots, \beta_{\tilde{s}_0}$ are the active components: (i) $\alpha_1, \alpha_2$; (ii) $\beta_1, \beta_2$; (iii) $\alpha_1, \ldots, \alpha_{\tilde{s}_0}$; (iv)$\beta_1, \ldots, \beta_{\tilde{s}_0}$; (v) $\alpha_1, \alpha_{\tilde{s}_0+1}, \beta_1, \beta_{\tilde{s}_0+1}$; (vi) $\alpha_1, \ldots, \alpha_p$; (vii) $\beta_1, \ldots, \beta_q$; and (viii) $\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q$.
Figure 3 depicts an example of such two-dimensional 95% confidence regions as obtained from

the simulations. We present the results of our simulation study in Table 4. Again, we observe far better coverage for the desparsified estimator and, as expected, correlation makes matters harder.

## 7 Analysis of the gene expression data

We apply our method on the gene expression data available as the dataset *eyedata* from the R package *flare* (Li et al., 2014). This dataset contains expression levels of 200 genes from 120 rats and is extracted from the study of Scheetz et al. (2006). We refer to this paper for more details. The response variable is the expression level of the TRIM32 gene, known to be linked to the Bardet-Biedl syndrome. Scientists are interested in finding the genes whose expressions are highly correlated with that of gene TRIM32. In the pre-analysis, we observe one observation which could be considered as an outlier and decide to remove it from the dataset for our analysis. We standardize the data and observe an average correlation of 0.5 between the 200 covariates. Our dataset is a subset of the one analyzed by Huang et al. (2010) who estimated every component separately as a spline function in an additive model and observed that most of the covariates are highly non-linear. This motivates us to introduce many variables in the nonlinear part of our model. The decision of which variables should go into the nonlinear part $\eta(Z^\top\alpha)$ and which ones should be in the linear part $X^\top\beta$ is not always straightforward. The split can for example be done based on the particular meaning of the variables (see Xia and Härdle (2006)) or based on the scatterplot of each variable versus the response (see Xia et al. (1999) and Zhang et al. (2012)). We decide to use the second strategy: we fit a semiparametric regression model for each variable separately, compute the degrees of freedom of the fits and choose 1.5 as threshold. Thus, variables which show a nonlinear behaviour (degrees of freedom larger than 1.5) go into the nonlinear part of our model and other ones go into the linear part. This leads to a dimension of 91 for $\alpha$ and 109 for $\beta$.

We fit a partially-linear single-index model $Y = \eta(Z^\top\alpha) + X^\top\beta$ to the data and choose the bandwidth by generalized cross-validation. We use the scaled Lasso to fit this high-dimensional model and obtain an active set containing 16 variables, including 5 components of $\alpha$. The fit of the semiparametric part of the model is shown in Figure 4, showing a clear need of the nonparametrically estimated function $\eta$ as opposed to using an identity function as in the linear model. The corresponding value of R-squared is 0.72. We apply our method for constructing confidence intervals and find 11 significant variables at a 5% significance level, of which 5 variables are components of $\alpha$, see Table 5.

Figure 5(a)-(b) show the 200 univariate confidence intervals and highlight some of them. From

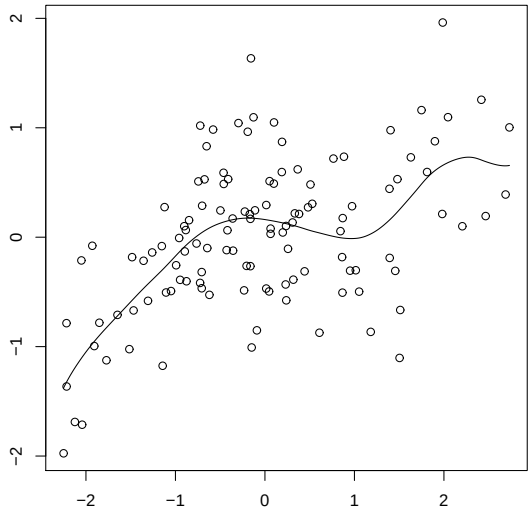


Figure 4: Scatterplot of the estimate $Z^\top \hat{\alpha}$ of the index versus $Y - X^\top \hat{\beta}$. The solid line is the estimate $\hat{\eta}(Z^\top \hat{\alpha}, \hat{\xi})$ of $\eta(Z^\top \alpha)$.

Figure 5(a) we can observe two opposite cases: $\alpha_{16}$ (VAR27) is selected by the Lasso but is not significant at a 95% confidence level, while $\alpha_{58}$ (VAR120) is set to 0 by the Lasso, but after correction, is shown to be significantly different from zero. The main advantage of our methodology is precisely this ability to take the uncertainty associated with the selection of the zero components into account. In Figure 5(b) we highlight such a variable selected by the Lasso and significant at the 5% level. The same kind of observations can be drawn from Table 5.

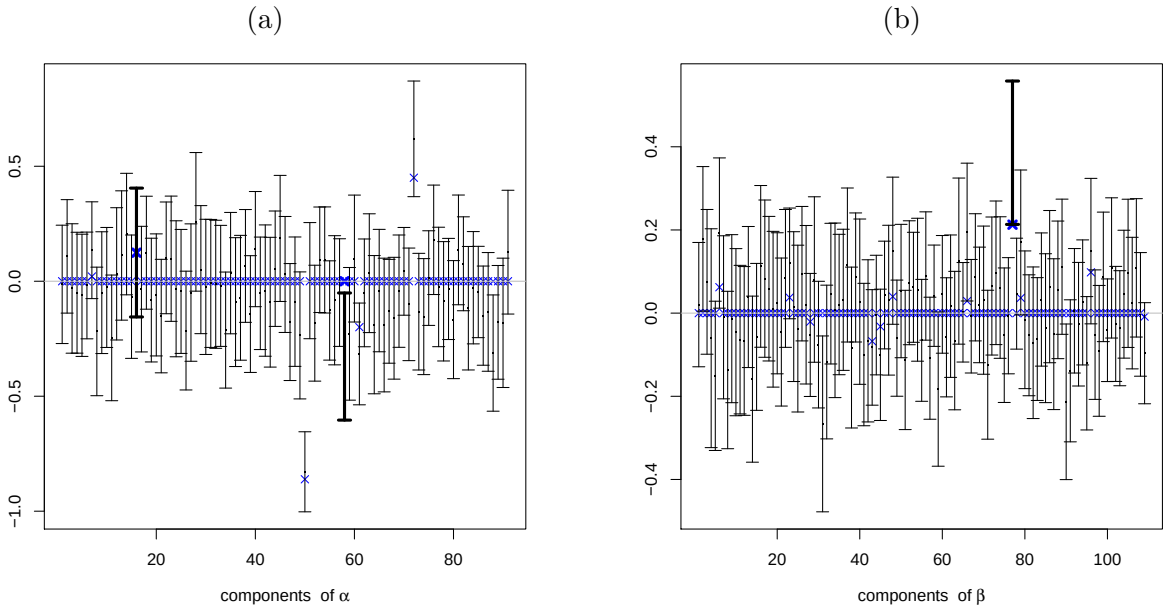


Figure 5: Penalized estimator and 95% univariate confidence intervals for the parameters of the PLSIM. (a) Components of $\alpha$ and (b) components of $\beta$. Some cases are highlighted and discussed in the text: $\alpha_{16}$ (VAR27), $\alpha_{58}$ (VAR120) and $\beta_{77}$ (VAR 153).

# 8 Discussion

We have shown that the desparsifying procedure has its benefits too for the partially linear single-index models. In general, we would recommend using the here proposed confidence intervals and their hypothesis testing equivalents, rather than working with the active sets of selected coefficients only. The simulations and data analysis have illustrated that the set of coefficients set to zero cannot be completely ignored as some of those values might better not have been set to zero. The consistency of selection is an asymptotic property, and comes with selection uncertainty as any model selection method does.

While we concentrated on models with a single index, the method readily extends to multiple indices of the form $\eta(Z_1^\top \alpha_1, \ldots, Z_d^\top \alpha_d)$. The single-index model takes $d = 1$. Due to the curse of dimensionality, it would be advised to take reasonably low values of $d$ (such as two or three) for efficient estimation.

It is of interest to extend these techniques to generalized partially linear single-index or multiple index models (see Carroll et al. (1997)). It is expected that a combination of the presented theory of this paper combined with the methods of Section 3 of van de Geer et al. (2014) leads to the expected results.

The steps we took to study the effect of a nonparametrically estimated function in the partially linear single-index model, is expected to open the way towards desparsifying methods in other nonparametric and semiparametric models as well. Such issues are beyond the scope of the current paper.

# 9 Proofs

We define $R(\alpha, \beta) = nP_n\dot{\rho}_\xi$ and use the same penalty parameter $\lambda$ for both model parts to result in a penalty function $p_\lambda(\alpha, \beta)$. With more notational complexity, the same steps of the proof hold for different penalty parameters for $\alpha$ and $\beta$. The penalized estimator is $\hat{\xi} = (\hat{\alpha}, \hat{\beta}) = \arg\min\{Q(\alpha, \beta) + p_\lambda(\alpha, \beta)\}$ and in this notation the desparsified estimator is $\hat{\xi}^{\text{desp}} = \hat{\xi} + n^{-1}\hat{\Theta}R(\hat{\alpha}, \hat{\beta})$. We use the following lemma's.

**Lemma 1.** *Under the conditions (C1) to (C6)(ii) we have*

$$\sup_{z,\alpha} \left| \hat{\eta}(z^\top \alpha; \alpha, \hat{\beta}) - \eta(z^\top \alpha_0) \right| = O_P\{(nh/\ln n)^{-1/2}\}, \tag{10}$$

$$\sup_{z,\alpha} \left| \hat{\eta}'(z^\top \alpha; \alpha, \hat{\beta}) - \eta'(z^\top \alpha_0) \right| = O_P\{(nh_1^3/\ln n)^{-1/2}\}. \tag{11}$$

*Proof of Lemma 1.* We first show (10). Using lemma A.4 from Wang et al. (2010), it holds that

$$\sup_{z,\alpha} \left| \hat{\eta}(z^\top \alpha; \alpha, \beta_0) - \eta(z^\top \alpha_0) \right| = O_P\{(nh/\ln n)^{-1/2}\}.$$

Hence we only need to show that

$$\sup_{z,\alpha} |\hat{\eta}(z^\top \alpha; \alpha, \hat{\beta}) - \hat{\eta}(z^\top \alpha; \alpha, \beta_0)| = O_P\{(nh/\ln n)^{-1/2}\}.$$

Using the equivalent kernel notation for local linear estimators (see Fan and Gijbels (1996), Sec. 3.2.2) with kernel functions $W_{n,j}$ it suffices to show that

$$\sup_{z,\alpha} |\sum_{j=1}^{n} W_{nj}(Z^\top \alpha, \alpha) X_j^\top (\hat{\beta} - \beta_0)| = O_P\{(nh/\ln n)^{-1/2}\}.$$

Using (C4) and (C6)(ii), and since $\sum_{j=1}^n W_{nj}(Z^\top \alpha, \alpha) X_j^\top (\hat{\beta} - \beta_0)$ is the local linear estimator using the one-dimensional observations $X_j^\top (\hat{\beta} - \beta_0)$, $j = 1, \ldots, n$, it satisfies the known rates for one-dimensional smoothing, which proves (10). The second part of the lemma for the estimation of the derivative curve can be shown similarly by using that $nh_1^5$ is bounded. □

**Lemma 2.** *Under the conditions of Theorem 1 we have for each* $j \in \{1, \ldots, p+q\}$

$$\sum_{i=1}^{n} \hat{\Theta}_{j,1}^{\top} Z_i \{\hat{\eta}'(Z_i^{\top}\hat{\alpha}; \hat{\xi}) - \eta'(Z_i^{\top}\alpha_0)\} = o_P(\sqrt{n}).$$

*where* $\hat{\Theta}_{j,1}^{\top}$ *is the vector formed by the first* $p$ *components of* $\hat{\Theta}_{j\bullet}$.

*Proof of lemma 2.* Similarly to (A.22) of Wang et al. (2010), we can show that

$$\sum_{i=1}^{n} \hat{\Theta}_{j,1}^{\top} Z_i \{\hat{\eta}'(Z_i^{\top}\hat{\alpha}; \hat{\alpha}, \beta_0) - \eta'(Z_i^{\top}\alpha_0)\} = o_P(\sqrt{n}).$$

It remains to show that $\sum_{i=1}^{n} \hat{\Theta}_{j,1}^{\top} Z_i \{\hat{\eta}'(Z_i^{\top}\hat{\alpha}; \hat{\alpha}, \beta_0) - \hat{\eta}'(Z_i^{\top}\hat{\alpha}; \hat{\alpha}, \hat{\beta})\} = o_P(\sqrt{n})$ or equivalently that $\sum_{i=1}^{n} \hat{\Theta}_{j,1}^{\top} Z_i \sum_{k=1}^{n} \widetilde{W}_{nk}(Z_i^{\top}\alpha, \alpha) X_k^{\top}(\hat{\beta} - \beta_0) = o_P(\sqrt{n})$, with $\widetilde{W}_{nk}$ the equivalent kernel notation for the estimation of the derivative curve. Using condition (C7), the convergence in distribution of the one-dimensional estimator $X^{\top}(\hat{\beta} - \beta_0)$ and the moment condition on the equivalent kernel (see Fan and Gijbels (1996), eq. (3.12)), the lemma is proven. □

*Proof of theorem 1.* We rewrite

$$\begin{aligned}
\hat{\xi}^{\text{desp}} - \xi_0 &= \hat{\xi} - \xi_0 + \tfrac{1}{n}\hat{\Theta} \textstyle\sum_{i=1}^{n} \{Y_i - \hat{\eta}(Z_i^{\top}\hat{\alpha}; \hat{\alpha}, \hat{\beta}) - X_i^{\top}\hat{\beta}\} \begin{bmatrix} Z_i \hat{\eta}'(Z_i^{\top}\hat{\alpha}; \hat{\alpha}, \hat{\beta}) \\ X_i \end{bmatrix} \\
&= \hat{\xi} - \xi_0 + \tfrac{1}{n}\hat{\Theta}\{R_1(\hat{\alpha}, \hat{\beta}) + R_2(\hat{\alpha}, \hat{\beta}) - R_3(\hat{\alpha}, \hat{\beta}) - R_4(\hat{\alpha}, \hat{\beta}) - R_5(\hat{\alpha}, \hat{\beta}) - R_6(\hat{\alpha}, \hat{\beta})\}
\end{aligned}$$

with

$$\begin{aligned}
R_1(\hat{\alpha}, \hat{\beta}) &= \sum_{i=1}^{n} \begin{bmatrix} \tilde{Z}_i \eta'(Z_i^{\top}\alpha_0) \\ \tilde{X}_i \end{bmatrix} \epsilon_i, \\
R_2(\hat{\alpha}, \hat{\beta}) &= \sum_{i=1}^{n} \begin{bmatrix} Z_i\{\hat{\eta}'(Z_i^{\top}\hat{\alpha}, \hat{\alpha}, \hat{\beta}) - \eta'(Z_i^{\top}\alpha_0)\} \\ 0 \end{bmatrix} \epsilon_i, \\
R_3(\hat{\alpha}, \hat{\beta}) &= \sum_{i=1}^{n} \begin{bmatrix} Z_i \eta'(Z_i^{\top}\alpha_0) \\ X_i \end{bmatrix} \{\hat{\eta}(Z_i^{\top}\hat{\alpha}, \hat{\alpha}, \hat{\beta}) - \hat{\eta}(Z_i^{\top}\alpha_0, \alpha_0, \beta_0) + X_i^{\top}(\hat{\beta} - \beta_0)\}, \\
R_4(\hat{\alpha}, \hat{\beta}) &= \sum_{i=1}^{n} \left[ \begin{bmatrix} Z_i \eta'(Z_i^{\top}\alpha_0) \\ X_i \end{bmatrix} \left\{\hat{\eta}(Z_i^{\top}\alpha_0, \alpha_0, \beta_0) - \eta(Z_i^{\top}\alpha_0)\right\} - \begin{bmatrix} \mathrm{E}[Z|Z_i^{\top}\alpha_0]\eta'(Z_i^{\top}\alpha_0) \\ \mathrm{E}[X_i|Z_i^{\top}\alpha_0] \end{bmatrix} \epsilon_i \right], \\
R_5(\hat{\alpha}, \hat{\beta}) &= \sum_{i=1}^{n} \begin{bmatrix} Z_i\{\hat{\eta}(Z_i^{\top}\hat{\alpha}, \hat{\alpha}, \hat{\beta}) - \eta(Z_i^{\top}\alpha_0)\}\{\hat{\eta}'(Z_i^{\top}\hat{\alpha}, \hat{\alpha}, \hat{\beta}) - \eta'(Z_i^{\top}\alpha_0)\} \\ 0 \end{bmatrix}, \\
R_6(\hat{\alpha}, \hat{\beta}) &= \sum_{i=1}^{n} \begin{bmatrix} Z_i\{\hat{\eta}'(Z_i^{\top}\hat{\alpha}, \hat{\alpha}, \hat{\beta}) - \eta'(Z_i^{\top}\alpha_0)\} \\ 0 \end{bmatrix} X_i^{\top}(\hat{\beta} - \beta_0).
\end{aligned}$$

We show that (i) $\hat{\Theta} R_k(\hat{\alpha}, \hat{\beta}) = o_P(\sqrt{n})$ for $k \in \{2, 4, 5\}$ and that

$$n(\hat{\xi} - \xi_0) - \hat{\Theta} R_3(\hat{\alpha}, \hat{\beta}) - \hat{\Theta} R_6(\hat{\alpha}, \hat{\beta}) = o_P(\sqrt{n}). \tag{12}$$

This will imply that $\sqrt{n}(\hat{\xi}^{\text{desp}} - \xi_0) = \frac{1}{\sqrt{n}}\hat{\Theta}\sum_{i=1}^{n}\begin{bmatrix}\tilde{Z}_i\eta'(Z_i^\top\alpha_0)\\ \tilde{X}_i\end{bmatrix}\epsilon_i + o_P(1)$ and prove the first part of the theorem.

Using Lemma 2, it is straightforward to show that $\hat{\Theta}R_2(\hat{\alpha},\hat{\beta})$ is $o_P(\sqrt{n})$. Using Lemma 1 and conditions (C4) and (C7), we have that

$$\hat{\Theta}_j^\top R_5(\hat{\alpha},\hat{\beta}) \le c\{nh/\ln(n)\}^{-1/2}\{nh_1^3/\ln(n)\}^{-1/2} = o_P(\sqrt{n}),$$

for each component $j$, which implies that $\hat{\Theta}R_5(\hat{\alpha},\hat{\beta}) = o_P(\sqrt{n})$.

To show (12), we write

$$\begin{aligned}
n(\hat{\xi}_j - \xi_j^0) - \hat{\Theta}_j^\top R_3(\hat{\alpha},\hat{\beta}) - \hat{\Theta}_j^\top R_6(\hat{\alpha},\hat{\beta}) &= n\left[e_j^\top - \hat{\Theta}_j^\top\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}Z_i\eta'(Z_i^\top\alpha_0)\\ X_i\end{bmatrix}^{\otimes 2}\right](\hat{\xi}-\xi_0)\\
&-\hat{\Theta}_j^\top\sum_{i=1}^{n}\begin{bmatrix}Z_i\eta'(Z_i^\top\alpha_0)\\ X_i\end{bmatrix}\{\hat{\eta}(Z_i^\top\hat{\alpha},\hat{\alpha},\hat{\beta}) - \hat{\eta}(Z_i^\top\hat{\alpha},\hat{\alpha},\beta_0)\}\\
&-\hat{\Theta}_j^\top\sum_{i=1}^{n}\begin{bmatrix}Z_i\eta'(Z_i^\top\alpha_0)\\ X_i\end{bmatrix}\{\hat{\eta}'(Z_i^\top\bar{\alpha},\bar{\alpha},\beta_0) - \eta'(Z_i^\top\alpha_0)\}Z_i^\top(\hat{\alpha}-\alpha_0) - \hat{\Theta}_j^\top R_6(\hat{\alpha},\hat{\beta})\\
= \;& R_{31} - R_{32} - R_{33} - \hat{\Theta}_j^\top R_6(\hat{\alpha},\hat{\beta}).
\end{aligned}$$

We treat $R_{31}$ and $\hat{\Theta}_j^\top R_6(\hat{\alpha},\hat{\beta})$ together. Using conditions (C6)(iii), (C8) and (C9), we have that

$$n\left(e_j^\top - \hat{\Theta}_j^\top\frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix}Z_i\hat{\eta}'(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta})\\ X_i\end{bmatrix}^{\otimes 2}\right)(\hat{\xi}-\xi_0) = o_P(\sqrt{n}).$$

It is thus sufficient to show that

$$\hat{\Theta}_j^\top\sum_{i=1}^{n}\left(\begin{bmatrix}Z_i\hat{\eta}'(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta})\\ X_i\end{bmatrix}^{\otimes 2} - \begin{bmatrix}Z_i\eta'(Z_i^\top\alpha_0)\\ X_i\end{bmatrix}^{\otimes 2}\right)(\hat{\xi}-\xi_0) - \hat{\Theta}_j^\top R_6(\hat{\alpha},\hat{\beta}) = o_P(\sqrt{n}) \qquad (13)$$

to prove that $R_{31} - \hat{\Theta}_j^\top R_6(\hat{\alpha},\hat{\beta}) = o_P(\sqrt{n})$. Writing $\hat{\Theta}_j^\top = (\hat{\Theta}_{j,1}^\top, \hat{\Theta}_{j,2}^\top)$ and $\hat{\xi}-\xi_0 = \begin{bmatrix}\hat{\alpha}-\alpha_0\\ \hat{\beta}-\beta_0\end{bmatrix}$, (13) can be decomposed as

$$\begin{aligned}
&\hat{\Theta}_{j,1}^\top\sum_{i=1}^{n}Z_i\{\hat{\eta}'(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta})^2 - \eta'(Z_i^\top\alpha_0)^2\}Z_i^\top(\hat{\alpha}-\alpha_0)\\
+\;& \hat{\Theta}_{j,2}^\top\sum_{i=1}^{n}X_i\{\hat{\eta}'(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta}) - \eta'(Z_i^\top\alpha_0)\}Z_i^\top(\hat{\alpha}-\alpha_0)\\
+\;& \hat{\Theta}_{j,1}^\top\sum_{i=1}^{n}Z_i\{\hat{\eta}'(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta}) - \eta'(Z_i^\top\alpha_0)\}X_i^\top(\hat{\beta}-\beta_0)\\
-\;& \hat{\Theta}_{j,1}^\top\sum_{i=1}^{n}Z_i\{\hat{\eta}'(Z_i^\top\hat{\alpha};\hat{\alpha},\hat{\beta}) - \eta'(Z_i^\top\alpha_0)\}X_i^\top(\hat{\beta}-\beta_0).
\end{aligned}$$

The last two elements sum up to zero and the first two elements are both $o_P(\sqrt{n})$ by Lemma 1 and condition (C6)(i). We thus have $R_{31} - \hat{\Theta}_j^\top R_6(\hat{\alpha}, \hat{\beta}) = o_P(\sqrt{n})$. Showing that $R_{32}$ and $R_{33}$ are $o_P(\sqrt{n})$ will imply that $n(\hat{\xi} - \xi_0) - \hat{\Theta}_j^\top R_3(\hat{\alpha}, \hat{\beta}) - \hat{\Theta}_j^\top R_6(\hat{\alpha}, \hat{\beta}) = o_P(\sqrt{n})$.
For $R_{32}$, using $\hat{\eta}(t; \alpha, \beta) = \sum_{j=1}^n W_{nj}(t, \alpha)(Y_j - X_j^\top \beta)$, we obtain

$$R_{32} = \hat{\Theta}_j^\top \sum_{i=1}^n \begin{bmatrix} Z_i \eta'(Z_i^\top \alpha_0) \\ X_i \end{bmatrix} \sum_{j=1}^n W_{nj}(Z_i^\top \hat{\alpha}, \hat{\alpha}) X_j^\top (\hat{\beta} - \beta_0).$$

Now, given that $\sum_{j=1}^n W_{nj}(Z_i^\top \hat{\alpha}, \hat{\alpha}) X_j^\top (\hat{\beta} - \beta_0) = X_i^\top (\hat{\beta} - \beta_0) + O(h^2)$, we have $R_{32} = o(n^{1/2}) + O(nh^2) = o(n^{1/2})$ by conditions (C4), (C6)(ii) and (C7).
Showing that $R_{33}$ is $o_P(\sqrt{n})$ is straightforward using Lemma 1, condition (C6)(ii) and the fact that $\bar{\alpha}$ is between $\alpha_0$ and $\hat{\alpha}$.
It remains now to show that $\hat{\Theta} R_4(\hat{\alpha}, \hat{\beta}) = o_P(\sqrt{n})$ to obtain the asymptotic distribution of $\sqrt{n}(\hat{\xi}^{\text{desp}} - \xi_0)$.
We write $\hat{\Theta} R_4(\hat{\alpha}, \hat{\beta}) = R_{4a} + R_{4b}$ with

$$\begin{aligned} R_{4a} &= \sum_{i=1}^n \hat{\Theta}_{j,1}^\top \eta'(Z_i^\top \alpha_0) \left[ Z_i \{ \hat{\eta}(Z_i^\top \alpha_0, \alpha_0, \beta_0) - \eta(Z_i^\top \alpha_0) \} - \mathrm{E}[Z | Z_i^\top \alpha_0] \epsilon_i \right], \\ R_{4b} &= \sum_{i=1}^n \hat{\Theta}_{j,2}^\top \left[ X_i \{ \hat{\eta}(Z_i^\top \alpha_0, \alpha_0, \beta_0) - \eta(Z_i^\top \alpha_0) \} - \mathrm{E}[X | Z_i^\top \alpha_0] \epsilon_i \right]. \end{aligned}$$

We only show that $R_{4a} = o_P(\sqrt{n})$, the treatment of $R_{4b}$ is similar. We prove that the mean and the variance of $n^{-1/2} R_{4a}$ tend to 0. Using $\mathrm{E}[\hat{\eta}(Z_i^\top \alpha_0, \alpha_0, \beta_0) - \eta(Z_i^\top \alpha_0)] = O(h^2)$ and the conditions (C2) and (C7), we have $\mathrm{E}[n^{-1/2} R_{4a}] \leq n^{-1/2} n \| \hat{\Theta}_{j,1}^\top Z \|_\infty \sup |\eta'(Z_i \alpha_0)| c h^2 \leq c n^{1/2} h^2$ which tends to 0 by condition (C4).
Regarding the variance of $n^{-1/2} R_{4a}$, using condition (C7) and lemmas A.2 and A.3 of Wang et al. (2010), we obtain

$$\begin{aligned} n^{-1} \mathrm{E}[R_{4a}^2] &\leq c n^{-1} \sum_{i=1}^n E\Big[ \Big\{ \hat{\Theta}_{j,1}^\top \{ \sum_{k=1}^n W_{ni}(Z_k^\top \alpha_0, \alpha_0) \eta'(Z_k^\top \alpha_0) Z_k - \eta'(Z_i^\top \alpha_0) \mathrm{E}[Z | Z_i^\top \alpha_0] \} \Big\}^2 \Big] \\ &\quad + \sum_{i=1}^n E\Big[ (\hat{\Theta}_{j,1}^\top Z_i)^2 \eta'(Z_i^\top \alpha_0)^2 \{ \sum_{j=1}^n W_{nj}(Z_i^\top \alpha_0, \alpha_0) \eta(Z_j^\top \alpha_0) - \eta(Z_i^\top \alpha_0) \Big)^2 \} \\ &\leq c(nh)^{-1} + c\sqrt{h} + cnh^4, \end{aligned}$$

which tends to 0 by (C4) and Lemma A.2 of Wang et al. (2010). This proves that $\hat{\Theta} R_4(\hat{\alpha}, \hat{\beta}) = o_P(\sqrt{n})$. Further, it follows that

$$\sqrt{n}(\hat{\xi}^{\text{desp}} - \xi_0) = \frac{1}{\sqrt{n}} \hat{\Theta} \sum_{i=1}^n \begin{bmatrix} \tilde{Z}_i \eta'(Z_i^\top \alpha_0) \\ \tilde{X}_i \end{bmatrix} \epsilon_i + o_p(1).$$

Using Lemma 1 and condition (C7), it is straightforward to show that we have a consistent estimator of the variance of $\sqrt{n}(\hat{\xi}^{\text{desp}} - \xi_0)$. Thus,

$$\hat{\Theta}_j^\top \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \tilde{Z}_i \eta'(Z_i^\top \alpha_0) \\ \tilde{X}_i \end{bmatrix}^{\otimes 2} \hat{\Theta}_k - \hat{\Theta}_j^\top \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} \tilde{Z}_i \hat{\eta}'(Z_i^\top \hat{\alpha}; \hat{\alpha}, \hat{\beta}) \\ \tilde{X}_i \end{bmatrix}^{\otimes 2} \hat{\Theta}_k = o(1).$$

□

## Acknowledgements

The authors thank the reviewers for their constructive comments. We acknowledge the support of the Fund for Scientific Research Flanders, KU Leuven grant GOA/12/14 and of the IAP Research Network P7/06 of the Belgian Science Policy. The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Hercules Foundation and the Flemish Government - department EWI.

## References

Carroll, R. J., Fan, J., Gijbels, I., Wand, M. P., 1997. Generalized partially linear single-index models. Journal of the American Statistical Association 92 (438), 477–489.
URL http://dx.doi.org/10.1080/01621459.1997.10474001

Chatterjee, A., Lahiri, S. N., 06 2013. Rates of convergence of the adaptive lasso estimators to the oracle distribution and higher order refinements by the bootstrap. Ann. Statist. 41 (3), 1232–1259.
URL http://dx.doi.org/10.1214/13-AOS1106

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 04 2004. Least angle regression. Ann. Statist. 32 (2), 407–499.
URL http://dx.doi.org/10.1214/009053604000000067

Fan, J., Gijbels, I., 1996. Local Polynomial Modelling and its Applications. Vol. 66. CRC Press.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96 (456), 1348–1360.
URL http://dx.doi.org/10.1198/016214501753382273

Hjort, N. L., Claeskens, G., 2003. Frequentist model average estimators. Journal of the American Statistical Association 98, 879–899.
URL http://dx.doi.org/10.1198/016214502388618861

Horowitz, J. L., 1998. Semiparametric Methods in Econometrics. Springer.

Huang, J., Horowitz, J. L., Wei, F., 08 2010. Variable selection in nonparametric additive models. Ann. Statist. 38 (4), 2282–2313.
URL http://dx.doi.org/10.1214/09-AOS781

Huang, J., Ma, S., Zhang, C.-H., 2008. Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica 18 (4), 1603.
URL http://www3.stat.sinica.edu.tw/statistica/j18n4/j18n420/j18n420.html

Javanmard, A., Montanari, A., 2014. Confidence intervals and hypothesis testing for high-dimensional regression. Journal of Machine Learning Research 15 (1), 2869–2909.
URL http://dl.acm.org/citation.cfm?id=2697057

Kabaila, P., 2009. The coverage properties of confidence regions after model selection. International Statistical Review 77 (3), 405–414.
URL http://dx.doi.org/10.1111/j.1751-5823.2009.00089.x

Leeb, H., Pötscher, B. M., Ewald, K., 05 2015. On various confidence intervals post-model-selection. Statistical Science 30 (2), 216–227.
URL http://dx.doi.org/10.1214/14-STS507

Li, X., Zhao, T., Wang, L., Yuan, X., Liu, H., 2014. flare: Family of Lasso Regression. R package version 1.2.0.
URL http://CRAN.R-project.org/package=flare

Liang, H., Liu, X., Li, R., Tsai, C.-L., 12 2010. Estimation and testing for partially linear single-index models. Ann. Statist. 38 (6), 3811–3836.
URL http://dx.doi.org/10.1214/10-AOS835

Lockhart, R., Taylor, J., Tibshirani, R. J., Tibshirani, R., 04 2014. A significance test for the lasso. The Annals of Statistics 42 (2), 413–468.
URL http://dx.doi.org/10.1214/13-AOS1175

Lu, J., Kolar, M., Liu, H., 2015. Post-regularization confidence bands for high dimensional nonparametric models with local sparsity. arXiv preprint arXiv:1503.02978.
URL http://arxiv.org/pdf/1503.02978.pdf

Ma, Y., Zhu, L., 2013. Doubly robust and efficient estimators for heteroscedastic partially linear single-index models allowing high dimensional covariates. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75 (2), 305–322.
URL http://dx.doi.org/10.1111/j.1467-9868.2012.01040.x

Meinshausen, N., Buhlmann, P., 06 2006. High-dimensional graphs and variable selection with the lasso. Ann. Statist. 34 (3), 1436–1462.
URL http://dx.doi.org/10.1214/009053606000000281

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L.,

Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., Stone, E. M., 2006. Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences 103 (39), 14429–14434.
URL http://www.pnas.org/content/103/39/14429.abstract

Scott, D. W., 2009. Multivariate Density Estimation: Theory, Practice, and Visualization. Vol. 383. John Wiley & Sons.

Sun, T., Zhang, C.-H., 2012. Scaled sparse linear regression. Biometrika.
URL http://dx.doi.org/10.1093/biomet/ass043

Tibshirani, R., 1996. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological) 58 (1), 267–288.
URL http://dx.doi.org/10.2307/2346178

van de Geer, S., Bhlmann, P., Ritov, Y., Dezeure, R., 06 2014. On asymptotically optimal confidence regions and tests for high-dimensional models. Ann. Statist. 42 (3), 1166–1202.
URL http://dx.doi.org/10.1214/14-AOS1221

Waldorp, L., 2015. Testing for graph differences using the desparsified lasso in high-dimensional data. Technical report, University of Amsterdam.

Wang, J.-L., Xue, L., Zhu, L., Chong, Y. S., 2010. Estimation for a partial-linear single-index model. The Annals of Statistics 38 (1), 246–274.
URL http://dx.doi.org/10.1214/09-AOS712

Wang, Q., Wu, R., 2013. Shrinkage estimation of partially linear single-index models. Statistics & Probability Letters 83 (10), 2324 – 2331.
URL http://dx.doi.org/10.1016/j.spl.2013.06.019

Xia, Y., Härdle, W., 2006. Semi-parametric estimation of partially linear single-index models. Journal of Multivariate Analysis 97 (5), 1162 – 1184.
URL http://dx.doi.org/10.1016/j.jmva.2005.11.005

Xia, Y., Tong, H., Li, W. K., 1999. On extended partially linear single-index models. Biometrika 86 (4), 831–842.
URL http://dx.doi.org/10.1093/biomet/86.4.831

Yu, Y., Ruppert, D., 2002. Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association 97 (460), 1042–1054.
URL "http://dx.doi.org/10.1198/016214502388618861"

Zhang, C.-H., Zhang, S. S., 2014. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76 (1), 217–242.
URL http://dx.doi.org/10.1111/rssb.12026

Zhang, J., Wang, T., Zhu, L., Liang, H., 2012. A dimension reduction based approach for estimation and variable selection in partially linear single-index models with high-dimensional covariates. Electron. J. Statist. 6, 2235–2273.
URL http://dx.doi.org/10.1214/12-EJS744

Zhu, L., Xue, L., 2006. Empirical likelihood confidence regions in a partially linear single-index model. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (3), 549–570.
URL http://dx.doi.org/10.1111/j.1467-9868.2006.00556.x

Zou, H., 2006. The adaptive lasso and its oracle properties. Journal of the American Statistical Association 101 (476), 1418–1429.
URL http://dx.doi.org/10.1198/016214506000000735

| | $n = 100$ | | | $n = 200$ | | | $n = 500$ | | |
|---|---|---|---|---|---|---|---|---|---|
| Avg cov | Unpen. | Penal. | Desp. | Unpen. | Penal. | Desp. | Unpen. | Penal. | Desp. |
| $p, q = 10$ | 0.79 | n/a | 0.92 | 0.81 | n/a | 0.91 | 0.86 | n/a | 0.92 |
| $C_{0,\alpha}$ | 0.70 | 0.60 | 0.84 | 0.66 | 0.60 | 0.83 | 0.74 | 0.71 | 0.90 |
| $C^c_{0,\alpha}$ | 0.73 | n/a | 0.90 | 0.74 | n/a | 0.88 | 0.79 | n/a | 0.90 |
| $C_{0,\beta}$ | 0.83 | 0.42 | 0.92 | 0.87 | 0.53 | 0.92 | 0.92 | 0.59 | 0.94 |
| $C^c_{0,\beta}$ | 0.86 | n/a | 0.96 | 0.90 | n/a | 0.95 | 0.93 | n/a | 0.95 |
| $p, q = 50$ | n/a | n/a | 0.94 | 0.92 | n/a | 0.95 | 0.90 | n/a | 0.94 |
| $C_{0,\alpha}$ | n/a | 0.66 | 0.64 | 0.85 | 0.72 | 0.94 | 0.84 | 0.74 | 0.92 |
| $C^c_{0,\alpha}$ | n/a | n/a | 0.93 | 0.90 | n/a | 0.95 | 0.87 | n/a | 0.93 |
| $C_{0,\beta}$ | n/a | 0.23 | 0.81 | 0.94 | 0.29 | 0.95 | 0.92 | 0.38 | 0.94 |
| $C^c_{0,\beta}$ | n/a | n/a | 0.96 | 0.94 | n/a | 0.96 | 0.94 | n/a | 0.95 |
| $p, q = 200$ | n/a | n/a | 0.96 | n/a | n/a | 0.96 | 0.99 | n/a | 0.95 |
| $C_{0,\alpha}$ | n/a | 0.68 | 0.59 | n/a | 0.75 | 0.67 | 0.84 | 0.79 | 0.96 |
| $C^c_{0,\alpha}$ | n/a | n/a | 0.96 | n/a | n/a | 0.95 | 0.99 | n/a | 0.95 |
| $C_{0,\beta}$ | n/a | 0.14 | 0.76 | n/a | 0.12 | 0.82 | 0.99 | 0.19 | 0.96 |
| $C^c_{0,\beta}$ | n/a | n/a | 0.97 | n/a | n/a | 0.97 | 0.99 | n/a | 0.96 |
| Avg length | | | | | | | | | |
| $p, q = 10$ | 0.12 | n/a | 0.11 | 0.07 | n/a | 0.07 | 0.04 | n/a | 0.04 |
| $C_{0,\alpha}$ | 0.07 | 0.06 | 0.07 | 0.04 | 0.03 | 0.04 | 0.02 | 0.02 | 0.02 |
| $C^c_{0,\alpha}$ | 0.09 | n/a | 0.08 | 0.04 | n/a | 0.04 | 0.02 | n/a | 0.02 |
| $C_{0,\beta}$ | 0.16 | 0.14 | 0.15 | 0.09 | 0.09 | 0.09 | 0.05 | 0.05 | 0.05 |
| $C^c_{0,\beta}$ | 0.16 | n/a | 0.15 | 0.09 | n/a | 0.09 | 0.05 | n/a | 0.05 |
| $p, q = 50$ | n/a | n/a | 0.14 | 0.13 | n/a | 0.11 | 0.04 | n/a | 0.04 |
| $C_{0,\alpha}$ | n/a | 0.09 | 0.09 | 0.09 | 0.05 | 0.08 | 0.02 | 0.02 | 0.02 |
| $C^c_{0,\alpha}$ | n/a | n/a | 0.10 | 0.10 | n/a | 0.09 | 0.03 | n/a | 0.03 |
| $C_{0,\beta}$ | n/a | 0.18 | 0.17 | 0.16 | 0.10 | 0.14 | 0.06 | 0.05 | 0.06 |
| $C^c_{0,\beta}$ | n/a | n/a | 0.17 | 0.16 | n/a | 0.14 | 0.06 | n/a | 0.06 |
| $p, q = 200$ | n/a | n/a | 0.21 | n/a | n/a | 0.07 | 0.17 | n/a | 0.11 |
| $C_{0,\alpha}$ | n/a | 0.17 | 0.16 | n/a | 0.05 | 0.04 | 0.11 | 0.03 | 0.08 |
| $C^c_{0,\alpha}$ | n/a | n/a | 0.18 | n/a | n/a | 0.04 | 0.12 | n/a | 0.09 |
| $C_{0,\beta}$ | n/a | 0.27 | 0.25 | n/a | 0.10 | 0.10 | 0.21 | 0.06 | 0.13 |
| $C^c_{0,\beta}$ | n/a | n/a | 0.24 | n/a | n/a | 0.10 | 0.21 | n/a | 0.13 |

Table 1: Simulation study. Average coverage (top) and length (bottom) of confidence intervals for nominal coverage of 0.95. Sparsity $\tilde{s}_0 = 2$, identity covariance matrix, $\sigma_\epsilon = 0.3$. The scaled lasso is used to compute the penalized estimator.

| | Independent | | | Toeplitz | | | Equicorrelated | | |
|---|---|---|---|---|---|---|---|---|---|
| | Unpen | Penal. | Desp. | Unpen. | Penal. | Desp. | Unpen. | Penal. | Desp. |
| Avg cov, $\sigma_\epsilon = 0.3$ | 0.81 | n/a | 0.93 | 0.75 | n/a | 0.96 | 0.69 | 0.65 | 0.93 |
| $C_{0,\alpha}$ | 0.71 | 0.64 | 0.87 | 0.51 | 0.41 | 0.83 | 0.56 | 0.42 | 0.64 |
| $C^c_{0,\alpha}$ | 0.74 | n/a | 0.91 | 0.65 | n/a | 0.95 | 0.57 | 0.54 | 0.91 |
| $C_{0,\beta}$ | 0.88 | 0.32 | 0.95 | 0.86 | 0.67 | 0.98 | 0.81 | 0.42 | 0.97 |
| $C^c_{0,\beta}$ | 0.88 | n/a | 0.96 | 0.86 | n/a | 0.98 | 0.82 | 0.82 | 0.98 |
| $\sigma_\epsilon = 1$ | 0.9 | n/a | 0.95 | 0.86 | n/a | 0.96 | 0.83 | 0.86 | 0.94 |
| $C_{0,\alpha}$ | 0.88 | 0.9 | 0.95 | 0.64 | 0.62 | 0.9 | 0.73 | 0.7 | 0.84 |
| $C^c_{0,\alpha}$ | 0.88 | n/a | 0.95 | 0.81 | n/a | 0.96 | 0.77 | 0.81 | 0.94 |
| $C_{0,\beta}$ | 0.91 | 0.44 | 0.95 | 0.92 | 0.85 | 0.96 | 0.9 | 0.66 | 0.95 |
| $C^c_{0,\beta}$ | 0.92 | n/a | 0.95 | 0.92 | n/a | 0.96 | 0.9 | 0.95 | 0.95 |
| Avg length, $\sigma_\epsilon = 0.3$ | 0.04 | n/a | 0.04 | 0.26 | n/a | 0.25 | 0.16 | 0.12 | 0.14 |
| $C_{0,\alpha}$ | 0.03 | 0.02 | 0.03 | 0.09 | 0.08 | 0.1 | 0.06 | 0.05 | 0.06 |
| $C^c_{0,\alpha}$ | 0.03 | n/a | 0.03 | 0.1 | n/a | 0.1 | 0.06 | 0.05 | 0.06 |
| $C_{0,\beta}$ | 0.06 | 0.06 | 0.06 | 0.42 | 0.34 | 0.38 | 0.27 | 0.2 | 0.22 |
| $C^c_{0,\beta}$ | 0.06 | n/a | 0.06 | 0.44 | n/a | 0.4 | 0.27 | 0.2 | 0.22 |
| $\sigma_\epsilon = 1$ | 0.13 | n/a | 0.15 | 0.44 | n/a | 0.46 | 0.29 | 0.26 | 0.29 |
| $C_{0,\alpha}$ | 0.08 | 0.08 | 0.09 | 0.16 | 0.15 | 0.18 | 0.11 | 0.1 | 0.12 |
| $C^c_{0,\alpha}$ | 0.09 | n/a | 0.1 | 0.17 | n/a | 0.19 | 0.11 | 0.1 | 0.12 |
| $C_{0,\beta}$ | 0.18 | 0.18 | 0.19 | 0.69 | 0.62 | 0.7 | 0.48 | 0.42 | 0.46 |
| $C^c_{0,\beta}$ | 0.18 | n/a | 0.19 | 0.73 | n/a | 0.74 | 0.48 | 0.42 | 0.46 |

Table 2: Simulation study. Average coverage and length of confidence intervals for nominal coverage of 0.95 over 1000 realizations with $n = 500$, $p, q = 50$ and $\tilde{s}_0 = 5$.

| | Original | | | | Desparsified | | |
|---|---|---|---|---|---|---|---|
| Avg cov | Unpen | SCAD | Adapt.L | Scaled L | SCAD | Adapt.L | Scaled L |
| Sparsity $\tilde{s}_0 = 2$ | 0.86 | n/a | n/a | n/a | 0.94 | 0.94 | 0.93 |
| $C_{0,\alpha}$ | 0.79 | 0.66 | 0.66 | 0.72 | 0.92 | 0.92 | 0.91 |
| $C^c_{0,\alpha}$ | 0.83 | n/a | n/a | n/a | 0.94 | 0.94 | 0.92 |
| $C_{0,\beta}$ | 0.88 | 0.89 | 0.84 | 0.35 | 0.94 | 0.94 | 0.94 |
| $C^c_{0,\beta}$ | 0.9 | n/a | n/a | n/a | 0.95 | 0.95 | 0.95 |
| Sparsity $\tilde{s}_0 = 5$ | 0.81 | n/a | n/a | n/a | 0.94 | 0.93 | 0.93 |
| $C_{0,\alpha}$ | 0.71 | 0.56 | 0.57 | 0.64 | 0.89 | 0.88 | 0.87 |
| $C^c_{0,\alpha}$ | 0.74 | n/a | n/a | n/a | 0.93 | 0.93 | 0.91 |
| $C_{0,\beta}$ | 0.88 | 0.84 | 0.82 | 0.32 | 0.95 | 0.94 | 0.95 |
| $C^c_{0,\beta}$ | 0.88 | n/a | n/a | n/a | 0.95 | 0.95 | 0.96 |
| Sparsity $\tilde{s}_0 = 10$ | 0.78 | n/a | n/a | n/a | 0.94 | 0.93 | 0.94 |
| $C_{0,\alpha}$ | 0.69 | 0.56 | 0.57 | 0.62 | 0.89 | 0.88 | 0.89 |
| $C^c_{0,\alpha}$ | 0.69 | n/a | n/a | n/a | 0.93 | 0.93 | 0.92 |
| $C_{0,\beta}$ | 0.85 | 0.83 | 0.82 | 0.3 | 0.95 | 0.95 | 0.96 |
| $C^c_{0,\beta}$ | 0.86 | n/a | n/a | n/a | 0.96 | 0.95 | 0.97 |

Table 3: Average coverage of confidence intervals for nominal coverage of 0.95 for several estimators over 1000 realizations with $n = 500$ and $p, q = 50$.

| | | Design: independent vars | | | Toeplitz correlated vars | | |
|---|---|---|---|---|---|---|---|
| Set of components | Size | Unpen. | Pen. | Desp. | Unpen | Pen. | Desp. |
| $n = 500$, $p, q = 10$, $\tilde{s}_0 = 5$, $\sigma_\epsilon = 0.3$ (10 nonzero) | | | | | | | |
| $\alpha_1, \alpha_2$ | 2 | 0.50 | 0.48 | 0.81 | 0.17 | 0.15 | 0.50 |
| $\beta_1, \beta_2$ | 2 | 0.88 | 0.40 | 0.93 | 0.69 | 0.49 | 0.96 |
| $\alpha_1, \ldots, \alpha_{\tilde{s}_0}$ | 5 | 0.35 | 0.34 | 0.75 | 0.06 | 0.06 | 0.26 |
| $\beta_1, \ldots, \beta_{\tilde{s}_0}$ | 5 | 0.82 | 0.15 | 0.92 | 0.50 | 0.15 | 0.96 |
| $\alpha_1, \alpha_{\tilde{s}_0+1}, \beta_1, \beta_{\tilde{s}_0+1}$ | 4 | 0.54 | 0.42 | 0.84 | 0.19 | 0.14 | 0.72 |
| $\alpha_1, \ldots, \alpha_p$ | 10 | 0.19 | 0.09 | 0.65 | 0.01 | 0.01 | 0.12 |
| $\beta_1, \ldots, \beta_q$ | 10 | 0.78 | n/a | 0.93 | 0.34 | n/a | 0.96 |
| $\alpha_1, \ldots, \beta_q$ | 20 | 0.20 | n/a | 0.70 | 0.00 | n/a | 0.14 |
| $n = 500$, $p, q = 50$, $\tilde{s}_0 = 5$, $\sigma_\epsilon = 0.3$ (10 nonzero) | | | | | | | |
| $\alpha_1, \alpha_2$ | 2 | 0.60 | 0.51 | 0.85 | 0.21 | 0.18 | 0.60 |
| $\beta_1, \beta_2$ | 2 | 0.84 | 0.13 | 0.94 | 0.82 | 0.51 | 0.98 |
| $\alpha_1, \ldots, \alpha_{\tilde{s}_0}$ | 5 | 0.42 | 0.33 | 0.79 | 0.00 | 0.04 | 0.28 |
| $\beta_1, \ldots, \beta_{\tilde{s}_0}$ | 5 | 0.78 | 0.01 | 0.95 | 0.72 | 0.10 | 0.98 |
| $\alpha_1, \alpha_{\tilde{s}_0+1}, \beta_1, \beta_{\tilde{s}_0+1}$ | 4 | 0.60 | 0.47 | 0.89 | 0.40 | 0.29 | 0.89 |
| $\alpha_1, \ldots, \alpha_p$ | 50 | 0.00 | n/a | 0.42 | 0.00 | n/a | 0.10 |
| $\beta_1, \ldots, \beta_q$ | 50 | 0.29 | n/a | 0.97 | 0.27 | n/a | 0.99 |
| $\alpha_1, \ldots, \beta_q$ | 100 | 0.00 | n/a | 0.48 | 0.00 | n/a | 0.16 |
| $n = 500$, $p, q = 10$, $\tilde{s}_0 = 2$, $\sigma_\epsilon = 0.3$ (4 nonzero) | | | | | | | |
| $\alpha_1, \ldots, \alpha_{\tilde{s}_0}$ | 2 | 0.72 | 0.71 | 0.89 | 0.51 | 0.50 | 0.66 |
| $\beta_1, \ldots, \beta_{\tilde{s}_0}$ | 2 | 0.89 | 0.36 | 0.92 | 0.84 | 0.52 | 0.93 |
| $\alpha_1, \alpha_{\tilde{s}_0+1}, \beta_1, \beta_{\tilde{s}_0+1}$ | 4 | 0.65 | 0.46 | 0.83 | 0.39 | 0.35 | 0.76 |
| $\alpha_1, \ldots, \alpha_p$ | 10 | 0.40 | 0.22 | 0.75 | 0.10 | 0.00 | 0.30 |
| $\beta_1, \ldots, \beta_q$ | 10 | 0.84 | n/a | 0.93 | 0.71 | n/a | 0.94 |
| $\alpha_1, \ldots, \beta_q$ | 20 | 0.40 | n/a | 0.78 | 0.10 | n/a | 0.33 |
| $n = 500$, $p, q = 50$, $\tilde{s}_0 = 2$, $\sigma_\epsilon = 0.3$ (4 nonzero) | | | | | | | |
| $\alpha_1, \ldots, \alpha_{\tilde{s}_0}$ | 5 | 0.79 | 0.74 | 0.92 | 0.32 | 0.55 | 0.78 |
| $\beta_1, \ldots, \beta_{\tilde{s}_0}$ | 5 | 0.86 | 0.16 | 0.94 | 0.81 | 0.35 | 0.94 |
| $\alpha_1, \alpha_{\tilde{s}_0+1}, \beta_1, \beta_{\tilde{s}_0+1}$ | 4 | 0.68 | 0.33 | 0.90 | 0.48 | 0.57 | 0.84 |
| $\alpha_1, \ldots, \alpha_p$ | 50 | 0.09 | n/a | 0.58 | 0.02 | n/a | 0.38 |
| $\beta_1, \ldots, \beta_q$ | 50 | 0.47 | n/a | 0.94 | 0.26 | n/a | 0.97 |
| $\alpha_1, \ldots, \beta_q$ | 100 | 0.07 | n/a | 0.60 | 0.01 | n/a | 0.43 |

Table 4: Average coverage of multivariate confidence regions confidence intervals for nominal coverage of 0.95, with 500 observations, 20 or 100 variables with different sparsity.

**Partially linear single index model**:

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ **components**: | | | | | | | | | | | | | | |
| Variables | 13 | 27 | 102 | 120 | 124 | 140 | 195 | | | | | | | |
| Selected by scaled Lasso | × | × | × | | × | × | | | | | | | | |
| Significant at 5% | | | × | × | × | × | × | | | | | | | |
| $\beta$ **components**: | | | | | | | | | | | | | | |
| Variables | 3 | 11 | 54 | 62 | 66 | 87 | 90 | 96 | 134 | 153 | 157 | 171 | 180 | 200 |
| Selected by scaled Lasso | | × | × | × | | × | × | × | × | × | × | | × | × |
| Significant at 5% | × | × | | | × | | | | × | × | | × | | |

Table 5: Important variables for the partially linear single-index model fit. The parameter $\alpha$ is of dimension 91 and the parameter $\beta$ of dimension 109.

**FACULTY OF ECONOMICS AND BUSINESS**
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be